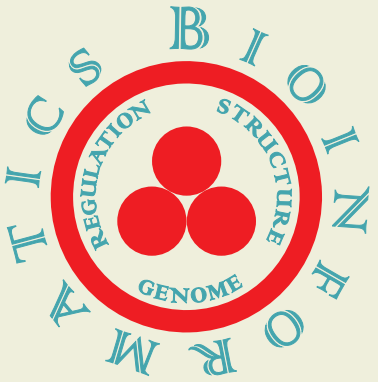
PROCEEDINGS OF THE FIFTH INTERNATIONAL CONFERENCE ON BIOINFORMATICS OF GENOME REGULATION AND STRUCTURE



VOLUME 3

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

RUSSIAN ACADEMY OF SCIENCES
SIBERIAN BRANCH

INSTITUTE OF CYTOLOGY AND GENETICS

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

Edited by
N. Kolchanov, R. Hofestädt

Volume 3

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

Novosibirsk
2006

INTERNATIONAL PROGRAM COMMITTEE

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

LOCAL ORGANIZING COMMITTEE

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

ISBN 5-7692-0848-1 (V.3)
ISBN 5-7692-0845-7

Our sponsors

Organizers

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

Institute of Cytology and Genetics, SB RAS

Siberian Branch of the Russian Academy of Sciences

Vavilov Society of Geneticists and Breeders

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

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

Grants

INTAS Grant

Russian Foundation for Basic Research

Contents

PART 5. COMPARATIVE AND EVOLUTIONARY GENOMICS AND PROTEOMICS

PART 6. OTHER TOPICS RELATED WITH BIOINFORMATICS

PART 7. SHORT ABSTRACTS

Introduction

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

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

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

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

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

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

PART 4. GENE NETWORKS THEORY: MATHEMATICAL PROBLEMS AND SOFTWARE

DEVELOPMENT OF A COMPUTER SYSTEM FOR THE AUTOMATED RECONSTRUCTION OF MOLECULAR-GENETIC INTERACTION NETWORKS

Aman E.E.[1], Demenkov P.S.[2], Pintus S.S.[3], Nemiatov A.I.[3], Apasieva N.V.[1], Dubovenko E.A.[1], Ignatieva E.V.[1], Podkolodny N.L.[1], Ivanisenko V.A.[*1]

[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia; [2] Institute of Mathematics, SB RAS, Novosibirsk, 630090, Russia; [3] Novosibirsk State University, Novosibirsk, 630090, Russia

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

Key words: information extraction, text mining, molecular-genetic interaction networks

SUMMARY

Motivation: The body of data related to molecular-genetic objects reported in the literature and stored in the computer databases keeps growing substantially over the years. The benefit of using computer assisted tools to overview and analyze bulky information is that it allows for the integration of information extracted from the databases with that retrieved from the automated processing of published reports.

Result: We are creating a system enabling us to detect interactions involving genes, proteins, low molecular weight substances, and human diseases. The broad basis is the information we retrieved from the databases about the interactions of these objects, protein and gene functional characteristics, gene expression profiles and also automated analysis of PubMed abstracts that relies on statistical approach. By using a combination of text- and data-mining, robust results can be achieved for the detection and systematization of information pertaining to molecular-genetic object interactions.

INTRODUCTION

Molecular biologists owe credit to the current achievements of computer-assisted reconstruction of gene networks. The respectable precision and higher recall of methods is, indeed impressive. The software GeneScene (Leroy, Chen, 2002) and MedScan (Daraselia *et al*., 2004) were developed to predict protein-protein and protein-gene interactions using text-mining. The software implements algorithms on the basis of an in-depth linguistic analysis of published texts. The authors of the systems have reported high precision, over 90 %, and recall of about 20 %. The systems based on linguistic text analysis yield very accurate estimates for the interaction detected in the text under consideration. However, their accuracy is restricted by the credibility of the analyzed paper. Low efficiency is another drawback.

Cooper have generated a simple system that predicts protein-protein interactions by using text-mining methods based on the search of particular words that describe the interactions, of protein name synonyms, and simple rules for their occurrence in the retrieved article. The system is admirably efficient, its accuracy is about 60 % (Cooper, Kershenbaum, 2005).

With this in mind, it appeared expedient to develop a combined approach that would enable us to use the shallow parsing of the PubMed abstracts. The envisaged combination will also include data on protein function, their intercellular location, gene expression profiles, among others, which can determine object interaction occurrence. The novelty of our approach is to use neural network algorithm for combining data obtained from abstracts with functional-structural data, which allow us to improve the prediction exactness. The accuracy of the automated reconstructed gene networks will be increased through a set of experimentally supported interactions retrieved from the GENE, KEGG and other publicly available databases. The system will give the user the opportunities to analyze the networks for interactions associated with biological processes and diseases of interest.

METHODS AND ALGORITHMS

We have developed a complex algorithm to detect real interactions between molecular-genetic objects. The algorithm was grounded on analysis of the PubMed abstracts and data on the functional-structural characteristics of the objects. The algorithm is schematically represented in Fig. 1.

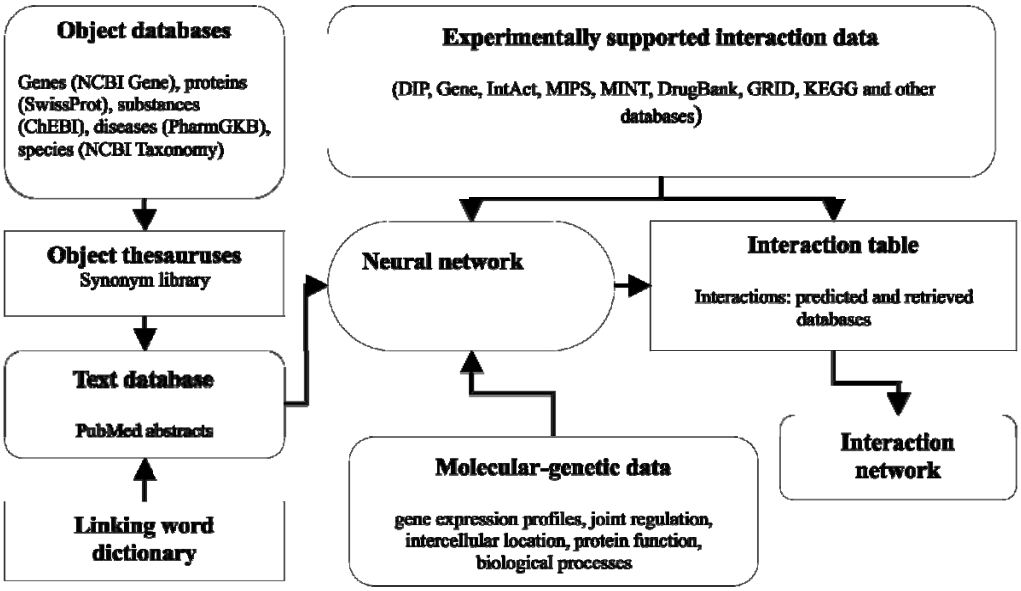


Figure 1. Schematic representation of the molecular-genetic interaction recognition algorithm.

Thesauruses (dictionaries of synonyms) were generated for databases of genes (NCBI GENE), proteins (SwissProt), substances (ChEBI), diseases (PharmGKB), and species (NCBI Taxonomy). Each entry in the dictionary contained a list of synonyms for each and every object, also references to the entries in the databases from which the information was retrieved.

To increase parsing accuracy of the abstracts, we listed the linking words whose occurrence in the sentence, along with the names of two objects, were evidences of interaction between them.

The collected dictionaries were used for parsing the PubMed abstracts. Four matrices for the names of objects and linking words occurrence in the texts were derived from the texts. Matrix building obeyed one of following rules:

- Two object names must occur in the same article's name.
- Two object names must occur in the same abstract.

- Two object names must occur in the same sentence.
- Two object names must occur in the same sentence and be joined by linking words.

A back propagation network (BPN) was used to predict object interactions. To train the neural network, experimentally supported data were retrieved from the DIP, Gene, IntAct, MIPS, MINT, DrugBank, GRID, KEGG and other relevant databases.

Besides the object name occurrence matrices, we exploited additional information about the objects. The information included expression profiles, gene coregulation, object location in the cell, protein functions, and certain biological processes. The idea of training of the neural network was to make it capable of answering the question: "Can two molecular-genetic objects really interact?"

Taken together, the results provided by the BPN and the experimentally supported data were used for building the matrices for the pairwise interactions. A fragment of the matrix derived interaction network can be reconstructed depending on the parameters chosen by the user. These parameters may include concrete objects, network size, network degree of relationship among others.

RESULTS

Dictionaries of object name synonyms. Dictionaries of object name synonyms were compiled from the SwissProt database. To list the synonyms, card fields ID, Synonyms were utilized. Synonym lists from two cards were joined when the synonym sets in two cards overlapped by more than 50 %. To compile the dictionary of gene name synonyms, we took advantage of the ID, Gene name, Gene description, and Gene aliases fields, also of the NCBI GENE database. The compilation of the dictionary of substance name synonyms was based on the ChEBI database data and its Name, IUPAC Name, Synonym fields. Information about substance hierarchy was also included in the dictionary. The dictionary of disease name synonyms was compiled from the information in the PharmGKB database using its Name, Alternate names, Related gene, Related drug fields. The two latter fields contain information about genes and substances (drugs), associated with diseases. To compile the dictionary of words that are synonyms of species names, we used the NCBI Taxonomy database, fields Name, GeneBank Common name, and Common name.

All the synonyms shorter than 3 characters and those overlapped with the authorized English dictionary (20,000 words) were omitted.

Table 1. The statistics for the dictionaries of synonyms

	Proteins	Genes	Diseases	Substances	Species
Number of entries	71025	1402525	4066	28339	219391
Terms (synonyms) per entry	1.991	2.071	9.004	1.858	1.317
Average length per term (words)	3.036	1.560	2.465	1.354	2.756
Ambiguous terms	7936 (6.292%)	146013 (6.651%)	3 (0.008%)	5305 (12.05%)	499 (0.173%)
Terms overlapping with English	169 (0.134%)	1952 (0.089%)	199 (0.544%)	170 (0.386%)	83 (0.028%)

The dictionary for the gene names contains the maximum entry names (Table 1, row 1). Diseases dictionary is the richest containing more than 9 terms per entry (Table 1, row 2). The compactness of Substances and Genes dictionaries is the highest, 1.35 and 1.56 words per term respectively; the least is Proteins dictionary with more than 3 words per term (Table 1, row 3). On the one hand long terms can be recognized more precisely due to lesser ambiguity, but on the other hand longer terms tend to vary significantly in

texts. The number of ambiguous terms, i.e. the number of those associated with multiple object IDs, is very large for Substances, it is smaller for Proteins and Genes, and still smaller for Diseases (Table 1, row 4). The percent of term overlapping with English dictionary in Diseases is maximum among all the dictionaries (exceeding 0.5 %), its minimum estimate is for Species (less than 0.03) (Table 1, row 5). The term ambiguity is a source of many recognition errors. To disambiguate terms in the same dictionary the text can be analyzed in order to find keywords corresponding to a definite object. To solve ambiguity of terms in different dictionaries we can search nouns like "gene" or "protein" and verbs "express" or "catalyze" determining object type close to term mention.

Linking word dictionary. This dictionary was compiled manually using expertise analysis of the PubMed abstracts. The dictionary contains more than 100 words and word-forms occurring in sentences that describe interactions between objects, for example "regulate expression" or "interact with".

The next steps of the algorithm are now under development. This process is ongoing.

ACKNOWLEDGEMENTS

Work was supported in part by Russian Foundation for Basic Research (Nos 04-01-00458, 05-04-49283, 06-04-49556), the CRDF Rup2-2629-NO-04, the State contract No. 02.434.11.3004 of 01.04.2005 and No. 02.467.11.1005 of 30.09.2005 with the Federal Agency for science and innovation "Identification of promising targets for the action of new drugs on the basis of gene network reconstruction" federal goal-oriented technical program "Study and development of priority directions in science and technique, 2002-2006, Interdisciplinary integrative project for basic research of the SB RAS No. 115 "Development of intellectual and informational technologies for the generation and analysis of knowledge to support basic research in the area of natural science".

REFERENCES

Cooper J.W., Kershenbaum A. (2005) Discovery of protein-protein interactions using a combination of linguistic, statistical and graphical information. *BMC Bioinformatics*, **6,** 143.

Daraselia N., Yuryev A., Egorov S., Novichkova S., Nikitin A., Mazo I. (2004) Extracting human protein interactions from MEDLINE using a full-sentence parser. *Bioinformatics*, **20,** 604–611.

Leroy G., Chen H. (2002) Filling preposition-based templates to capture information from medical abstracts. *Pac Symp. Biocomput.*, 350–361.

RECONSTRUCTION AND COMPUTER ANALYSIS OF THE FATTY ACID β-OXIDATION GENE NETWORK REGULATED BY THE PPAR TRANSCRIPTION FACTORS

Aman E.E.*[*], *Levitsky V.G., Ignatieva E.V.
Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia
[*] Corresponding author: e-mail: aman@bionet.nsc.ru

Key words: gene networks, fatty acid β-oxidation, recognition of transcription factors binding sites

SUMMARY

Motivation: The receptors of PPARs (Peroxisome proliferator-activated receptors) are the regulators of a wide range of processes related to the energy metabolism in humans and animals. The gene network regulating the β-oxidation of fatty acids may be valuable for developing technologies and approaches to computer-based analysis of gene networks regulated by the PPAR factors.

Results: The GeneNet technology was used to reconstruct the fatty acid β-oxidation gene network. The network was reconstructed on the basis of the data in the literature. The application of the computer methods enabled us to enhance the gene network by detecting new regulatory interactions. The detection of new genes, the potential targets for PPAR, was implemented by using SiteGA, the method for the recognition of the transcription factors binding sites.

INTRODUCTION

The PPAR transcription factors are proteins of the nuclear receptor superfamily that show a wide range of biological activity. The spectrum of ligands that regulate PPARs activity is very wide. They include both endogenous molecules (free fatty acids, eicosanoids, leucotrienes, prostaglandins, among others) and also numerous artificially synthesized substances. The distinguished isotypes are PPARα, PPARβ, and PPARγ. PPARα stimulates the transport of fatty acids (FA) from the intestine to liver and their further utilization through FA oxidation of different types. PPARγ is a potent modulator of the differentiation and function of adipose cells. PPARβ presumably plays an important role in the implantation and decidualization of the uterus. PPARγ is involved in inflammatory response (Desvergne, Wahli, 1999). The aim of this work was to reconstruct and analyze the gene network for the FA β-oxidation regulated by the PPAR transcription factors using computer-based approaches.

METHODS AND ALGORITHMS

The GeneNet technology was used to reconstruct the gene network for the FA β-oxidation (Ananko *et al.*, 2005). To recognize the binding sites for the PPAR transcription factors, we used the SiteGA method. It is based on the calculation of the

correlation between the dinucleotide frequencies in the transcription factor binding site (Levitsky *et al.*, 2004). A sample of binding sites from the Transcription Regulatory Regions Database (TRRD) was used for training (Kolchanov *et al.*, 2002). The sample was composed of 39 sites with a classical structure consisting of 2 hexanucleotides AGGTCA separated by a spacer of 1 nucleotide (Kliewer *et al.*, 1992). The accuracy of the recognition method was as follows: false negative (FN) 50 %, false positive (FP) $1.62*10^{-4}$. The promoter regions of 74 genes of the gene network of the FA β-oxidation were extracted through the BLAT system (The BLAST like alignment tool), using the data for mRNA sequences for the RefSeq database (Pruitt *et al.*, 2005).

RESULTS AND DISCUSSION

The gene network for FA β-oxidation. The gene network includes the data for human, mouse and rat and incorporates the main scheme for the β-oxidation of saturated FAs in the mitochondria and peroxisomes and also two subschemes that describe the β-oxidation of the unsaturated FAs. The total informational content of the gene networks is: 76 genes, 80 proteins, and 189 reactions. The main β-oxidation scheme describes the β-oxidation of saturated FAs in the mitochondria (Fig. 1). The saturated free fatty acids (FFAs) bind to coenzyme A (CoA) in the cytoplasm and transport to the mitochondria through the carnitin-dependent mechanism. Here occurs sequential shortening of the carbon chain of acyl-CoA through Acyl-CoA dehydrogenase (ACADM) and 2-enoyl-CoA hydratase/3-hydroxyacyl-CoA dehydrogenase/3-ketothiolase (ECHA/ECHB) associated with the formation of acetyl-CoA, which is then metabolized in the citric acid cycle. The FAs with a chain of more than 20 carbon residues are oxidized in the peroxisomes. The difference from the mitochondrial cycle is the formation of the H_2O_2 at the first step.

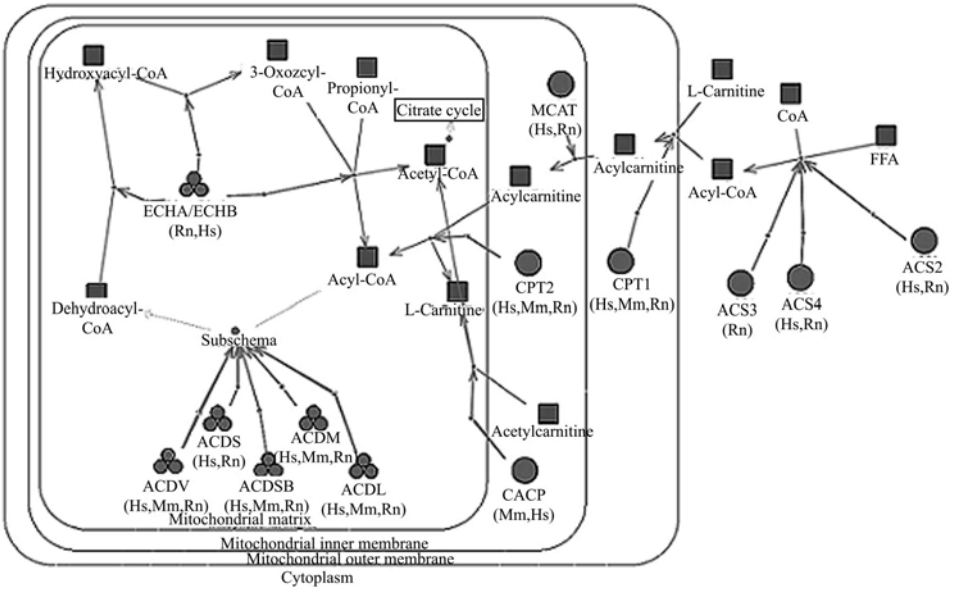


Figure 1. A fragment of the gene network for FA β-oxidation (β-oxidation of FAs in the mitochondria). Designations: Filled ovals, proteins; filled rectangles, substances; filled rectangles with arrows, genes; and arrows denote the regulatory effects and the reactions.

Oxidation of the unsaturated FAs includes besides the above listed steps, other additional. This oxidation can pass through several pathways, depending on the location of the double bond in the FA molecule.

There is experimental evidence that, of the 76 human, murine and rat gene of the FA β-oxidation gene network, transcription of 7 is regulated by the PPARs transcription factors (Fig. 3). Being the ligands of PPARs, FFAs can activate the PPAR/RXR heterodimer and enhance the expression of the PPAR regulated genes by binding to the ligand-binding domain of PPARs.

Contextual analysis of PPAR binding sites. Using the SiteGA method, we analyzed aligned sequences of 150 bp. and established correlations between dinucleotide frequencies (Fig. 2). The established correlations mean that the dinucleotides occupy specific locations that are of importance for DNA-protein interactions and that also define the dynamic DNA properties.

52 % of the significant correlations are in the region of the core site (at positions 71–83). Most of the correlations are at the positions that correspond to the end of the second hexanucleotide of the site (at positions 80–83) In the spacer region (positions 76–78), the number of correlations falls significantly. These data are consistent with the fact that the PPARs binding site has a two-core structure. The significant correlations were detected not only in the sequence of the core site, also beyond its confines. This may be taken to mean that the binding mechanism of PPARs to DNA is complex.

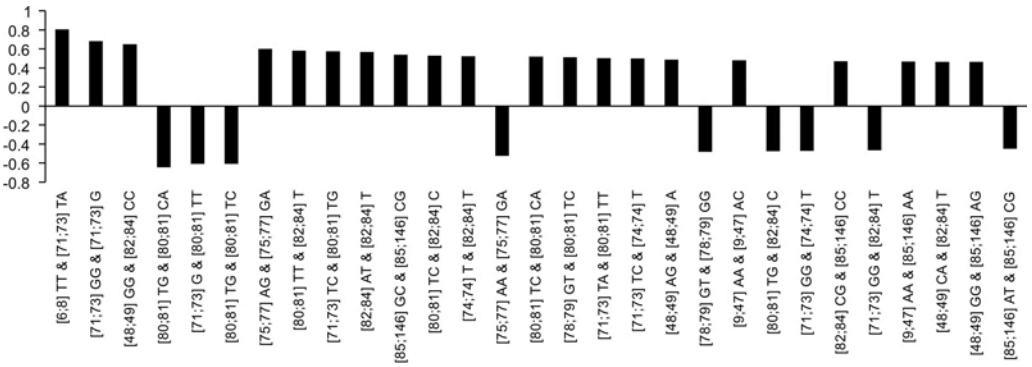


Figure 2. The diagrams of significant correlations used in the recognition method construction. The dinucleotide positions 71–83 corresponds to the core site.

Detection of new genes - the potential targets for the PPAR factors. Using the SiteGA method, we recognized PPAR binding sites in the regulatory regions (-1000/-1 relative to the transcription start) of 74 genes from the FA β-oxidation regulatory network. In all, 19 new potential binding sites were detected in 14 human, murine and rat genes (Table 1). On the basis of the *in silico* obtained results the gene network of FA β-oxidation was supplemented by 11 regulatory interactions (Fig. 3).

Table 1. PPARs binding sites recognized in the regulatory regions of genes of the network for the FA β-oxidation using the SiteGA method

Gene name	Protein name	Site direction*	Position**
ACADM (Hs)	acyl-CoA dehydrogenase (medium-chain)	→	-864
		←	-362
ACADSB (Hs)	acyl-CoA dehydrogenase (branched-chain)	→	-173
Acox1 (Rn)	acyl-CoA oxydase	→	-365
Acsl1 (Mm)	acyl-CoA synthetase 1	→	-294
Acsl3 (Rn)	acyl-CoA synthetase 3	→	-855
Cpt1a (Rn)	carnitine palmitoyltransferase 1a	→	-798

Gene name	Protein name	Site direction*	Position**
		→	-797
		→	-766
Cpt2 (Mm)	carnitine palmitoyltransferase 2	→	-362
CRAT (Hs)	carnitine acetyltransferase	→	-332
Crot (Mm)	carnitine octanoyltransferase	←	-620
Crot (Rn)	carnitine octanoyltransferase	←	-570
DECR1(Hs)	2,4-dienoyl-CoA reductase	←	-537
Ehhadh (Mm)	enoyl-CoA, hydratase/3-hydroxyacyl CoA dehydrogenase	←	-922
		←	-929
RXRA (Hs)	retinoid-X-receptor α	→	-741
SLC25A20 (Hs)	carnitine-acylcarnitine translocase	←	-840
		→	-325

* Direction: "→" - forward, "←" - backward; ** The positions are indicated relative to the transcription start.

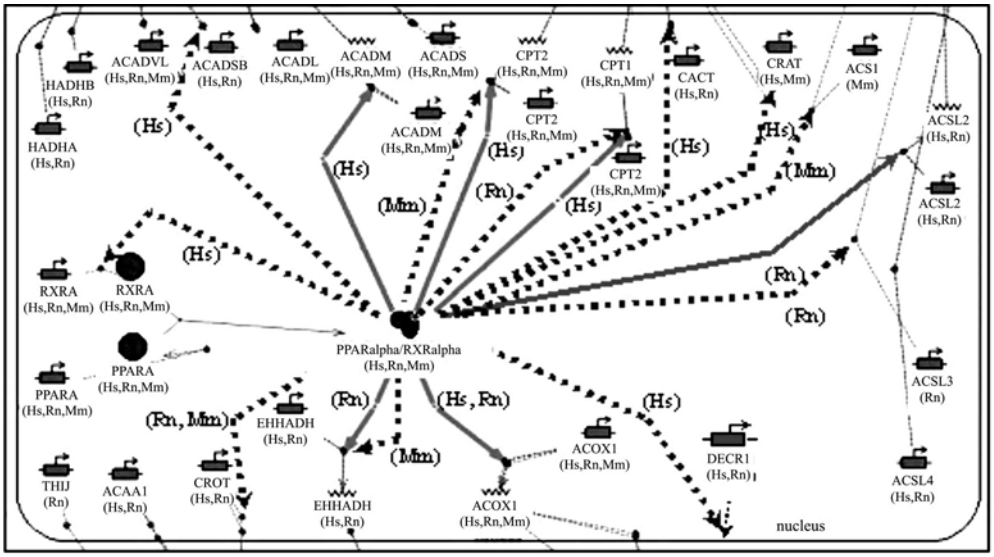


Figure 3. A fragment of the gene network for the β-oxidation of the FAs (regulation of the expression of the genes by the heterodimeric transcription factor PPARα/RXRα. Continuous arrows indicate the regulatory effects described in the literature and stored in the GeneNet database. Discontinuous arrows point to the regulatory events detected by the SiteGA method. The organisms (Hs-human, Rn-rat, Mm-mouse) are in parentheses.

ACKNOWLEDGEMENTS

The work was supported in part by Russian Foundation for Basic Research (No. 05-04-49111), State contract No. 02.434.11.3004 of 01.04.2005 and No. 02.467.11.1005 of 30.09.2005 with the Federal Agency for science and innovation; Interdisciplinary integrative project for basic research of the SB RAS No. 115 "Development of intellectual and informational technologies for the generation and analysis of knowledge to support basic research in the area of natural science".

REFERENCES

Ananko E.A. *et al.* (2005) GeneNet in 2005. *Nucl. Acids Res*., **33,** 425–427.

Desvergne B., Wahli, W. (1999) Peroxisome proliferator-activated receptors: nuclear control of metabolism. *Endocr. Rev.,* **20,** 649–688.

Kliewer S.A. *et al.* (1992) Convergence of 9-cis retinoic acid and peroxisome proliferator signalling pathways through heterodimer formation of their receptors. *Nature,* **358,** 771–774.

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

Levitsky V.G. *et al.* (2004) Analysis of the context features of SF-1 binding site and development of a criterion for SF-1 regulated gene recognition by the SiteGA method. *Proceeding of the Fourth Intern. Conf. on Bioinformatics of Genome Regulation and Structure-2004*, Novosibirsk, **1,** 119–122.

Pruitt K.D. *et al*. (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucl. Acids Res.*, **33**, 501–504.

OPTIMAL CONTROL TASKS IN TERMS OF THE GENE NETWORK MODELS

***Bezmaternykh K.D.*[*2], *Nikulichev Yu.V.*[1], *Likhoshvai V.A.*[2], *Matushkin Yu.G.*[2], *Latipov A.F.*[1], *Kolchanov N.A.*[2]**

[1] Institute of Theoretical and Applied Mechanics, SB RAS, Novosibirsk, 630090, Russia;
[2] Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia
[*] Corresponding author: e-mail: bezmate@bionet.nsc.ru

Key words: genes grid, stable statement, nonlinear programming, optimal control

SUMMARY

Motivation: Disfunction of gene networks caused by mutations in genes is frequently pathogenic. Optimal policies on a gene network can recover its normal function. Search for such policies requires the development of optimal control methods in dynamics of gene networks.

Results: A typical problem of optimal control of the functioning dynamics of gene networks was formulated. Using a new software suite, the optimal control problem on cholesterol gene network was resolved.

INTRODUCTION

The gene network consists of numerous elements with complex nonlinear links (Kolchanov *et al*., 2000). For this reason, correction of certain GN function parameters can have undesirable side effects (unacceptable values of other parameters). To recover the normal work of a damaged GN, it is necessary to find a control that minimizes side effects. All this gives rise to the optimal control problem in gene network models. Here, we formulate the control problem for the major GN versions (Kolchanov *et al*., 2000). Also, using the mathematical model of cholesterol biosynthesis in the cell, as an example, we resolve the issue of the recovery of the normal GN function damaged by mutation (Kolchanov *et al*., 2004).

METHODS AND ALGORITHMS

The tested object was the mathematical model of cholesterol biosynthesis in the cell. (Ratushny *et al*., 2003). To resolve the optimal control problem, we implemented the grid method (parameterization of the control functions) based on the optimization package "POISK" (Nikulichev, 2005). We used a variant of the penalty function method (Latipov, 1974) to take into account the constraints imposed on optimal control search.

RESULTS

Optimal control in GN: formulation of the problem. The mathematical model that describes the dynamics of the $\vec{w}$ component concentrations takes the form of a system of ordinary differential equations.

$$\begin{aligned} &\frac{d\vec{w}}{d\tau} = \vec{\varphi}\left(\vec{w}, \vec{a}, \vec{u}\right); \\ &\vec{u} \in D = \left\{ \vec{u} : \vec{u}^{min} \le \vec{u}\left(\tau\right) \le \vec{u}^{max} \right\} \end{aligned}, \qquad (1)$$

where $\vec{w}$ is the vector of the variable GN states, $\vec{a}$ are the coefficients that define the interaction of the system's components with each other and the environment: synthesis and degradation, rate of introduction of certain environmental components or their release. Change in one or several components of the $\vec{a}$ vector simulates mutation in the GN. A part of the parameters $\vec{a}$ form the control vector $\vec{u}(t)$, whose range of acceptable values is given by the D set. It is assumed that the conditions for the existence and the sole solution of the system (1) are satisfied.

The optimal control problem is formulated in general terms. Let the normal GN function in the time interval [0,T] be known. Let us define it as $\vec{w}_0(t)$. *The task:* in the [0,T] interval, using the control $\vec{u}(t)$, to reorganize the dynamics of the mutant GN function in such a way that it would maximally agree with the normal $\vec{w}_0(t)$ if the functional restrictions are imposed and the minimum condition of a functional F (energy, substance expenditures, etc.) is met.

A crucial parameter of treatment efficiency is its duration. Hence, the task of normal GN recovery will be subdivided as follows:

Task 1. Minimization of the time for the transition from the initial state of the mutant network $\vec{w}(0)$ to the range of the norm $\vec{w}_0(t)$

$$\begin{aligned} &\vec{w}(0) \Rightarrow^{\vec{u}(t)} \Rightarrow \vec{w}(\tau) \in B = \left\{ w_i(\tau) : \left| w_i(\tau) - w_{0i}(\tau) \right| \le \Delta_i, i = \overline{1,n} \right\}; \\ &\phi(\vec{w}, \vec{a}, \vec{u}) \le 0; \\ &\vec{u}(t) \in D = \left\{ \vec{u} : \vec{u}^{min} \le \vec{u}(t) \le \vec{u}^{max} \right\}; \; t \in [0, \tau]; \\ &\tau \Rightarrow \min_{\vec{u}(\tau)}, \tau \in [0, T] \end{aligned} \qquad (2)$$

The set B produces a corridor of acceptable values for the dynamic variables where the GN dynamics is regarded as normal. The vector-function ϕ is the continuous function on a set of variables, and it defines the functional constraints imposed on the process.

Task 2. To keep the GN dynamics in the range of the normal trajectory of the phase variables under the conditions of the minimum functional $F\left(\vec{w}(t)\right)$.

$$
\begin{aligned}
& w(t) \in B = \left\{ \mathrm{w}_{\mathrm{i}}(t) : \left| \mathrm{w}_{\mathrm{i}}(t) - \mathrm{w}_{0\,\mathrm{i}}(t) \right| \le \Delta_{\mathrm{i}}, \mathrm{i} = \overline{1,\mathrm{n}} \right\}; \\
& \phi(\vec{\mathrm{w}}, \vec{a}, \vec{\mathrm{u}}) \le 0; \\
& \vec{\mathrm{u}}(t) \in D = \left\{ \vec{\mathrm{u}} : \vec{\mathrm{u}}^{min} \le \vec{\mathrm{u}}(t) \le \vec{\mathrm{u}}^{max} \right\}; \\
& \mathrm{F}(\vec{\mathrm{w}}(t)) \Rightarrow \min_{\vec{\mathrm{u}}(t)}, \ \tau \in [0, T], t \in [\tau, T]
\end{aligned}
\tag{3}
$$

$\phi(\cdot)$, $F(\cdot)$ are the continuous functions on a set of variables. The vector-function ϕ defines the functional restrictions imposed on the process. The F functional defines expenditures of energy, essential substances and etc.

The optimal control problem. Recovery of the steady state in the model of cholesterol biosynthesis in the cell is provided as an illustrative example (Kolchanov *et al*., 2004).

1. Search of the acceptable control transferring the system from the mutant state to the range of the norm.
2. Minimization of the time for the transition τ (Task (1)).

At the first step, control was set in the class of constant functions in the entire interval of the effect; at the second step, the control functions varied in the $[0;\tau]$ interval; in the $[\tau;T]$ interval, the results obtained at the first step were accepted. The corridor for the control variables was of the ± 2 orders. Table 1 gives the values for the phase variables in the stationary state ($\vec{w}_0$, $\vec{w}_{0M}$, $\vec{w}_1$). The values for the control variables at both steps are tabulated below in Table 2.

Table 1. Steady state values for the phase variables

I	Phase variable name	Normal value $\vec{w}_0$	Acceptable normal range [%]	Mutant steady state point $\vec{w}_{0M}$	Control results (1) $\vec{w}_1$
1	SREBP1	$5.82 \cdot 10$	20	$8.74 \cdot 10$	$4.77 \cdot 10$
2	HMGR____	$5.50 \cdot 10^3$	20	$8.03 \cdot 10^3$	$6.05 \cdot 10^3$
3	cholin	$3.37 \cdot 10^5$	20	$2.24 \cdot 10^5$	$2.76 \cdot 10^5$
4	LDLR____	$7.12 \cdot 10^3$	20	$1.04 \cdot 10^4$	$8.43 \cdot 10^3$
5	LDL ____	$2.24 \cdot 10^5$	5	$1.19 \cdot 10^7$	$2.29 \cdot 10^2$
6	SRP prot	$5.94 \cdot 10^{-1}$	no constraints	$8.92 \cdot 10^{-1}$	$6.20 \cdot 10^{-1}$
7	SRP-chol	$4.00 \cdot 10^5$	20	$4.00 \cdot 10^5$	$3.42 \cdot 10^5$
8	inACAT	$1.01 \cdot 10^4$	20	$1.03 \cdot 10^4$	$9.13 \cdot 10^3$
9	ACAT____	$9.85 \cdot 10^3$	20	$9.78 \cdot 10^3$	$8.83 \cdot 10^3$
10	Cholestr	$1.69 \cdot 10^5$	20	$1.18 \cdot 10^5$	$1.96 \cdot 10^5$

Table 2. Control results

I	Control variable name	Value by default $\vec{u}_0$	Control (1) $\vec{u}_1$	Control (2) $\vec{u}_2$
1	Pcholin	$1.0 \cdot 10^5$	$8.294 \cdot 10^6$	$7.903 \cdot 10^6$
2	PLDL____	$5.0 \cdot 10^3$	$8.339 \cdot 10^1$	$6.869 \cdot 10^1$
3	PinACAT	$1.0 \cdot 10^4$	$8.981 \cdot 10^3$	$1.194 \cdot 10^4$
4	PSRPprot	1.0	$8.543 \cdot 10^{-1}$	$8.547 \cdot 10^{-1}$
5	HMGRgene	1.0	1.331	1.310
6	LDLRgene	1.0	1.432	1.022
7	PSREBP1	$1.0 \cdot 10^3$	$7.850 \cdot 10^2$	$7.861 \cdot 10^2$
8	Hydrolas	$1.0 \cdot 10^5$	$6.457 \cdot 10^4$	$1.219 \cdot 10^5$

At the second step, we succeeded in reducing the time of the transition process from $6.37 \cdot 10^5$ sec to $8.32 \cdot 10^4$ sec. Fig. 1 presents the graphs for the transition processes at both steps.

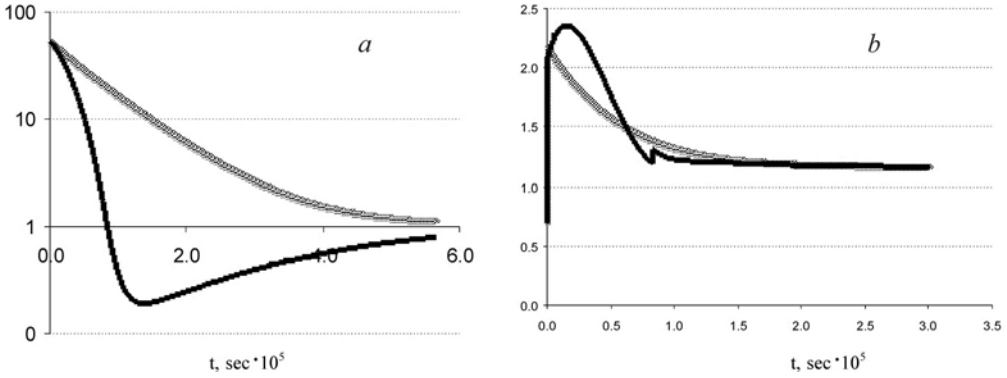


Figure 1. Results of the optimal control by the cholesterol biosynthesis model in the cell. Light curve denotes the system's transition from the mutant steady state to the range of the norm (single control interval). Dark curve represents the results of two-interval control: minimization the time of the transition ($t \in [0; 8.32.10^4]$) and confinement in the range of norm ($t \in [8.32.10^4, 6.0.10^5]$). Time (sec) is plotted along the X axis, the concentration related to the norm ($\vec{w}_0$) is plotted along the Y. Concentrations: low-density lipoproteins (logarithmic scale), aA; cholesterol ether, B.

DISCUSSION

A direct practical application of the recovery task of the dynamics of GN function is the improvement of realistic approaches to the treatment of diseases. This will provide information for the development of medicinal preparations. Another application of optimal control in GN is the design of experiments with *a priori* given constraints imposed on changes in known system's parameters to better understand other parameters.

It is hoped that the formulated task of the recovery of the normal GN dynamics is well applicable to any other gene networks no matter how complex its dynamics might be and parameters numerous. The goal is the search of an optimal control that would ensure the more rapid recovery of health. This is the hallmark feature of cure. In seeking computer short-cuts to therapy, the functional constrains imposed by biological systems on the recovery process cannot be discounted.

Currently, work designed to recover the stationary state of the expanded model (about 100 variables) of the genetic system controlling cholesterol homeostasis in mammalian cells subjected to mutations is underway.

ACKNOWLEDGEMENTS

Work was supported in part by Russian Foundation for Basic Research Nos 05-04-49283, 06-04-49556, the State contract No. 02.434.11.3004 of 01.04.2005 and No. 02.467.11.1005 of 30.09.2005 with the Federal Agency for science and innovation "Identification of promising targets for the action of new drugs on the basis of gene network reconstruction" federal goal-oriented technical program "Study and development of priority directions in science and technique, 2002–2006, Interdisciplinary integrative project for basic research of the SB RAS No. 115 "Development of intellectual and informational technologies for the generation and analysis of knowledge to support basic research in the area of natural science" and NSF:FIBR (Grant EF–0330786), Molecular-cellular biology and Evolutionary program. The authors are grateful to Aleksandr Ratushny for valuable discussions.

REFERENCES

Kolchanov N.A., Latypov A.F., Lihoshvaj V.A., Matushkin Yu.G., Nikulichev Yu.V., Ratushnyj A.V. (2004) A method of solving problems of optimal control in dynamics of gene networks. *Transactions of the Russ. Acad. Sci*., **6**, pp. 36–45.

Kolchanov N.A., Ananko E.A., Kolpakov F.A., Podkolodnaja O.A., Ignatjeva E.V., Goryachkovskay, Stepanenko I.L. (2000) Gene networks. *Mol. Biol.*, **34**, 533–544.

Latipov A.F. (1974) About solving extremum problems with limitations. *Izvestiya Sibirskogo Otdeleniya Akademii nauk SSSR*, *seriya tehnicheskih nauk,* **13**, 49–50.

Nikulichev J.V. (2005) Numerical methods on a basis hermitian splines for solving theCauchy problem for systems of the ordinary differential equations, approximation of curves and surfaces in problems of optimal control. The abstracts of the thesis for a doctor's degree, Novosibirsk.

Ratushny A.V., Likhoshvai V.A., Ignatieva E.V., Goryanin I.I., Kolchanov N.A. (2003) Resilience of cholesterol concentration to a wide range of mutations in the cell. *Complexus*, **1**, 142–148.

MATRIX PROCESS MODELLING: ON ONE CLASS OF INFINITE-ORDER SYSTEMS OF DIFFERENTIAL EQUATIONS AND ON DELAY DIFFERENTIAL EQUATIONS

***Demidenko G.V.*[*1], *Khropova Yu.E.*[2], *Kotova T.V.*[2]**
[1] Sobolev Institute of Mathematics, SB RAS, Novosibirsk, 630090, Russia; [2] Novosibirsk State University, Novosibirsk, 630090, Russia
[*] Corresponding author: e-mail: demidenk@math.nsc.ru

Key words: genetic systems, mathematical models, process, ordinary differential equations, delay differential equations, initial problem

SUMMARY

Motivation: In recent years, intensive study of gene networks, genetically controlled metabolic paths, signal transduction paths, and other complex genetic-molecular systems has been started. Presently, these studies are reaching a qualitatively new level due to wide use of microarray analysis, which makes it possible to reveal functions of many hundreds and even thousands of genes in a single experiment. Analysis of huge amounts of experimental data, which reflect complex processes in genetic-molecular systems, is impossible without efficient mathematical methods.

Results: In the present paper we consider a class of models describing matrix branching processes with an unrestrictedly increasing number of stages. We show that the last components of solutions of corresponding systems of ordinary differential equations tend to a solution of one delay differential equation.

INTRODUCTION

Study of gene networks is one of the main problems of systemic biology. For this purpose methods of numerical and theoretical mathematical modeling are extensively used. Modeling makes it possible to take into account synthesis of tens and hundreds of thousands of intermediate states of DNA, RNA, and proteins. However, from a mathematical standpoint, the consideration of intermediate stages of synthesis leads to systems of ordinary differential equations of a very high order. Thus, the *high dimensionality problem* arises when modeling gene networks.

As a rule, solving systems with a large number of differential equations, researchers try to reduce the problem to solving systems of substantially smaller orders. At the present time, there is a huge number of works devoted to various methods of reduction of orders. In the paper (Likhoshvai *et al.*, 2004), studying a model of multi-stage substance synthesis without branching

$$\begin{aligned}
&\frac{dx_1}{dt} = g(x_n) - \frac{n-1}{\tau} x_1, \\
&\frac{dx_i}{dt} = \frac{n-1}{\tau}(x_{i-1} - x_i), \qquad i = 2, ..., n-1, \\
&\frac{dx_n}{dt} = \frac{n-1}{\tau} x_{n-1} - \theta x_n, \qquad \tau > 0, \quad \theta \geq 0,
\end{aligned} \tag{1}$$

it was proposed and substantiated a new method to research systems of ordinary differential equations of sufficiently high orders. Keeping structures of systems, we proposed to increase unrestrictedly the number of equations of the systems and to consider solutions of limit systems.

Such an approach for the system of multi-stage substance synthesis without branching (1) for $n >> 1$ allows us to find approximately the end product, i.e., the component $x_n(t)$. As was shown in (Likhoshvai *et al.*, 2004), there is a close connection between solutions of the system (1) as $n \to \infty$ and solutions of the delay differential equation

$$\frac{d}{dt} y(t) = -\theta y(t) + g(y(t-\tau)), \quad t > \tau. \tag{2}$$

In particular, if we consider the Cauchy problem for (1) with the zero initial conditions and $g(z) \in C^1(R)$, then $x_n(t) \to y(t)$, $n \to \infty$, $t \in [0, T]$, where $y(t)$ is a solution of (2); moreover, $y(t) = 0$, $t \in [0, \tau]$. Consequently, to find approximately the end product we need not solve the Cauchy problem for systems of the form (1) of high orders. It suffices to solve an initial problem for one delay differential equation. Then $x_n(t) \approx y(t)$ as $n \to \infty$; moreover, we can write estimates for the approximation (see Likhoshvai *et al.*, 2004).

A development of this approach for constructing approximate solutions of ordinary differential equations of high orders can be found in the papers (Demidenko, Likhoshvai, 2005; Demidenko *et al.*, 2006).

In the present paper we point out a class of systems of ordinary differential equations

$$\frac{dx}{dt} = A_n x + F_n(t, x) \tag{3}$$

whose approximate solutions can be constructed by the method mentioned above. In particular, we establish a connection between these solutions and solutions of delay differential equations of the form (2). This class of systems includes the system of multi-stage substance synthesis with nonconstant rates of transition from the i th stage to the $i+1$ st stage.

RESULTS

Consider a series of systems of differential equations of the form (3). Each of the systems consists of n ordinary differential equations, linear terms are defined by an $n \times n$ numerical matrix A_n, and nonlinear terms are given by a vector-function $F_n(t, x)$ of the form

$$F_n(t, x) = (g(t, x_n), 0, ..., 0)^T.$$

We consider the series of systems of ordinary differential equations of the form (2) under the following conditions on the sequence $\{A_n\}$ of matrices and the function $g(t,z)$.

1. Let $\lambda_1^n,...,\lambda_n^n$ be the eigenvalues of the matrix A_n. Suppose that

$$\lambda_1^n \to -\theta, \quad n \to \infty, \operatorname{Re}\lambda_j^n \le -\frac{n-1}{\tau} + \lambda_0, \ j = 2,...,n,$$

where λ_0, $\theta \ge 0$, $\tau > 0$ are constant.

2. Assume that the cofactor α_n of the element $b_{1,n}$ of the matrix $(\lambda I - A) = (b_{i,j})$ does not depend on λ, and

$$|\alpha_n| \left(\frac{n-1}{\tau} \right)^{1-n} \le a < \infty.$$

3. Suppose that the convergence holds

$$\frac{1}{\alpha_n} \prod_{j=2}^{n} (\lambda_1^n - \lambda_j^n) \to e^{-\theta\tau}, \quad n \to \infty.$$

4. Assume that the function $g(t,z)$ is bounded and satisfies the Lipschitz condition

$$\sup_{t \ge 0} |g(t,z_1) - g(t,z_2)| \le L|z_1 - z_2|, \qquad z_1, z_2 \in R.$$

For each of the systems of the form (3) we consider the Cauchy problem with the zero initial conditions

$$\begin{aligned} &\frac{dx}{dt} = A_n x + F_n(t,x), \\ &x|_{t=0} = 0. \end{aligned} \tag{4}$$

We will increase unrestrictedly the number of equations of the system and consider only the last component of a solution of the Cauchy problem (4). Then we obtain a sequence $\{x_n(t)\}$ on the interval $[0,T]$.

Theorem 1. *The sequence $\{x_n(t)\}$ converges uniformly on the interval $[0,T]$, and the limit function $y(t)$ is a solution of the initial problem for a delay differential equation*

$$\begin{aligned} &\frac{d}{dt} y(t) = -\theta y(t) + g(t-\tau, y(t-\tau)), \qquad t > \tau, \\ &y(t) = 0, \qquad t \in [0,\tau]. \end{aligned}$$

We can show that an analog of Theorem 1 holds for a series of the Cauchy problem for systems of the form (3) with nonzero initial conditions. However, in this case we observe an interesting peculiarity that the sequence $\{x_n(t)\}$ converges in the $L_p(0,T)$ space, and the limit function $y(t)$ is a weak solution of a delay differential equation. We formulate this result in the case when the vector of initial conditions in the Cauchy problem (4) has a nonzero first component, while the others vanish, i.e.,

$$x|_{t=0} = x_0, \quad x_0 = (a,0,...,0)^T, \ a \ne 0.$$

Theorem 2. *For any $T > \tau$ the convergence holds*

$$\left\| x_n(t) - y(t), L_p(0,T) \right\| \to 0, \ p \ge 1, \qquad n \to \infty;$$

moreover, $y(t)$ *is a weak solution of the initial problem for a delay differential equation*

$$\frac{d}{dt}y(t) = -\theta y(t) + g(t-\tau, y(t-\tau)), \qquad t > \tau,$$

$$y(t) = 0, \qquad t \in [0, \tau), \qquad y(\tau + 0) = a.$$

Complete proofs of corresponding assertions can be found in (Demidenko *et al*., 2006).

DISCUSSION

From the proofs of the theorems in (Demidenko *et al*., 2006) we obtain uniform estimates of the difference module

$$\left|x_n(t) - y(t)\right| \le \beta(n), \quad t \in [\tau, T], \quad \beta(n) = O(n^{-1/4}), \quad n \to \infty. \tag{5}$$

Consequently, we have an effective method for constructing an approximation of the *n*th component $x_n(t)$ of the solution of the Cauchy problem (4) for $n >> 1$. Namely, to find approximately $x_n(t)$ we can solve the initial problem for the delay differential equation and use the estimate (5) in order to obtain accuracy of the approximation $x_n(t) \approx y(t)$, $n >> 1$. Obviously, using this method, we can study qualitative properties of the *n*th component $x_n(t)$.

ACKNOWLEDGEMENTS

The work was supported by the Russian Foundation for Basic Research (Nos 04-01-00458, 05-04-49068, 05-07-90274), the Siberian Branch of the Russian Academy of Sciences (Interdisciplinary integration project No. 24). The authors are grateful to Vitaly Likhoshvai for valuable discussions.

REFERENCES

Demidenko G.V., Likhoshvai V.A. (2005) On differential equations with retarded argument. *Sib. Mat. Zh*., **46**, 538–552; English transl. in *Sib. Math. J.*, **46**, 417–430.

Demidenko G.V., Likhoshvai V.A., Kotova T.V., Khropova Yu.E. (2006) On one class of systems of differential equations and on retarded equations. *Sib. Mat. Zh.*, **47**, 58-68; English transl. in *Sib. Math. J.*, **47**, 45-54.

Likhoshvai V.A., Fadeev S.I., Demidenko G.V., Matushkin Yu.G. (2004) Modelling of multi-stage synthesis without branching by an equation with delay. *Sib. Zh. Ind. Mat*., **7**, 73–94. (In Russ.).

MATRIX PROCESS MODELLING: ON A NEW METHOD OF APPROXIMATION OF SOLUTIONS OF DELAY DIFFERENTIAL EQUATIONS

***Demidenko G.V.*[*1], *Mudrov A.V.*[2]**
[1] Sobolev Institute of Mathematics, SB RAS, Novosibirsk, 630090, Russia; [2] Novosibirsk State University, Novosibirsk, 630090, Russia
[*] Corresponding author: e-mail: demidenk@math.nsc.ru

Key words: gene networks, mathematical models, ordinary differential equations, delay differential equations, initial problem

SUMMARY

Motivation: Matrix processes of replication, transcription, and translation, including hundreds and thousands of elongation stages of the same type, are essential part of natural and artificial genetic systems. To reflect adequately these processes in models of gene networks it is necessary to elaborate effective theoretical and numerical mathematical methods.

Results: In the present paper we solve the problem on approximation of solutions of delay differential equations by solutions of systems of ordinary differential equations. This result gives a new method in order to study biological processes described by means of systems of ordinary differential equations of very high orders; for example, multi-stage substance synthesis.

INTRODUCTION

In the present paper we continue to study connections between solutions of systems of a large number of ordinary differential equations

$$\begin{aligned}
&\frac{dx_1}{dt} = qmy_n - \frac{n-1}{\tau}x_1,\\
&\frac{dx_i}{dt} = \frac{n-1}{\tau}(x_{i-1} - x_i), \qquad i = 2,\dots,n-1,\\
&\frac{dx_n}{dt} = \frac{n-1}{\tau}x_{n-1} - mx_n,\\
&\frac{dy_n}{dt} = f(y_n, x_n), \quad 1 > q > 0, \quad \tau > 0, \quad m \in N,
\end{aligned} \tag{1}$$

and solutions of the delay differential equation

$$\frac{d}{dt}y(t) = f(y(t), qy(t-\tau)), \quad t > \tau. \tag{2}$$

We assume that the function $f(u,v)$ is bounded and satisfies the Lipschitz condition:
$\sup_{u,v\in R}|f(u,v)|= F<\infty$, $\quad |f(u_1,v_1)-f(u_2,v_2)|\le L_1|u_1-u_2|+L_2|v_1-v_2|$.

Let $y(t)$ be a solution of the equation (2), satisfying an initial condition

$$y(t)=\varphi(t),\quad 0\le t\le\tau, \tag{3}$$

where the function $\varphi(t)$ is a solution of the problem

$$\frac{d}{dt}\varphi(t)=f(\varphi(t),0),\quad \varphi|_{t=0}=y_0. \tag{4}$$

It was established (Demidenko, Likhoshvai, 2005) that $y(t)$ can be represented as the repeated limit

$$\lim_{m\to\infty}\lim_{n\to\infty}y_n^m(t)=y(t), \tag{5}$$

where $y_n^m(t)$ is the last component of a solution of the Cauchy problem for the system (1) with the initial conditions

$$x_1|_{t=0}=\ldots=x_n|_{t=0}=0,\quad y_n|_{t=0}=y_0. \tag{6}$$

Note that the representation (5) was proved in (Demidenko, Likhoshvai, 2005) on the interval $[0,t_0]$, where

$$\tau<t_0<\min\left\{\frac{1-q}{L_1},\frac{1}{L_2}\right\}.$$

Consequently, on "small" interval $[0,t_0]$ solutions of initial problems of the form (2), (3) for delay differential equations can be approximated by solutions of the Cauchy problems of the form (1), (6) for systems of a large number of ordinary differential equations:

$$y(t)\approx y_n^m(t),\quad m>>1,\quad n>>1. \tag{7}$$

In the present paper we strengthen our result (Demidenko, Likhoshvai, 2005) and prove that (5) holds on any interval $[0,T]$. We establish also uniform estimates for (7) and point out rate of the convergence (5).

RESULTS

We formulate the main results of the paper below.

Consider a series of the Cauchy problems of the form (1), (6) with a fixed m by unbounded increasing the number of equations n. Solving each problem and considering

only the last two components of solutions, we obtain a sequence of vector-functions $\{z_n^m(t)\}$, $z_n^m(t) = (x_n^m(t), y_n^m(t))$.

Theorem 1. *The sequence* $\{z_n^m(t)\}$ *converges uniformly as* $n \to \infty$ *on the interval* $[0, T]$

$$x_n^m(t) \to x^m(t), \quad y_n^m(t) \to y^m(t);$$

moreover, the limit vector-function $z^m(t) = (x^m(t), y^m(t))$ *is a solution of the system of integral equations*

$$x^m(t) = qm \int_0^{t-\tau} e^{-m(t-\tau-s)} \mathrm{y}^m(s) ds, \quad t > \tau, \tag{8}$$

$$y^m(t) = y_0 + \int_0^t f(y^m(s), x^m(s)) ds, \tag{9}$$

and $x^m(t) = 0$, $t \in [0, \tau]$.

Theorem 2. *The system of the integral equations (8), (9) has a unique solution* $z^m(t) = (x^m(t), y^m(t))$ *continuous on the interval* $[0, T]$, *and* $x^m(t) = 0$ *for* $t \in [0, \tau]$.
Obviously, the functions (8), (9) satisfy the system of the delay differential equations

$$\frac{d}{dt} x^m(t) = -m x^m(t) + qm y^m(t-\tau), \quad \frac{d}{dt} y^m(t) = f(y^m(t), x^m(t)). \tag{10}$$

Consider a sequence of the integral equations (8), (9) by unbounded increasing m. Solving each system of the form (8), (9), we obtain the sequence $\{z^m(t)\}$ of the vector-function $z^m(t) = (x^m(t), y^m(t))$.

Theorem 3. *The sequence* $\{z^m(t)\}$ *is convergent on the interval* $[\tau, T]$*:*

$x^m(t) \to x(t)$, $y^m(t) \to y(t)$, $m \to \infty$;

moreover,

$x(t) = qy(t-\tau)$,

$$y(t) = y_0 + \int_0^t f(y(s), 0) ds, \quad t \in [0, \tau],$$

$$y(t) = y_0 + \int_0^\tau f(y(s), 0) ds + \int_\tau^t f(y(s), qy(s-\tau)) ds, \; t \in [\tau, T].$$

From Theorem 3 we have

$y(t) \in C[0,T] \cap C^1(0,\tau) \cap C^1(\tau,T)$;

moreover, $y(t)$ is a solution of the problem (2), (3) for the delay differential equation with the initial function (4).

The next theorems give estimates for rate of the convergence (5).

Theorem 4. *For any fixed* $m \in N$, *the limit relations hold as* $n \to \infty$

$$\max_{t\in[0,T]} \left| x^m(t) - x_n^m(t) \right| = O(n^{-1/4}),$$

$$\max_{t\in[0,T]} \left| y^m(t) - y_n^m(t) \right| = O(n^{-1/4}).$$

Theorem 5. *The asymptotic equalities hold as* $m \to \infty$

$$\max_{t\in[j\tau+\delta,(j+1)\tau]} \left| x(t) - x^m(t) \right| = O\left(\frac{\ln m}{n} \right), \qquad \delta \in (0,\tau),$$

$$\max_{t\in[j\tau,(j+1)\tau]} \left| y(t) - y^m(t) \right| = O\left(\frac{\ln m}{n} \right)$$

where $[j\tau,(j+1)\tau] \subset [0,T]$.

DISCUSSION

Form Theorem 1–5 we obtain a method for approximation of solutions of delay differential equations of the form (2) with the conditions (3), (4) by using solutions of the Cauchy problems of the form (1), (6).

On the other hand, approximate finding the last two components of solutions of (1) with a sufficiently large number of differential equations and the initial conditions (6) can be obtained by means of solutions of the system of the integral equations (8), (9) or by using a corresponding initial problem for the system of two delay differential equations (10). In the case of sufficiently large m, approximate constructing the last component of a solution of (1) is reduced to solving the initial problem (2)–(4).

These results give a new method in order to study biological processes described by means of systems of ordinary differential equations of very high orders. Indeed, systems of high orders can be replaced by one or two delay differential equations. Thus, an analogous result was obtained in (Likhoshvai *et al*., 2004) for a system of differential equations modeling multi-stage substance synthesis without branching. As is known, synthesis of RNAs and proteins in gene networks involves a sufficiently large number (hundreds or even thousands) of intermediate stages. In this case, our results show that if the rate of each of the intermediate stages is sufficiently high, the kinetics of the end product output is practically independent of the kinetics if the internal synthesis stages. Everything depends on the regulatory mechanism of starting first stage of the synthesis and the delay time, which equals the average total time of all intermediate stages.

ACKNOWLEDGEMENTS

The work was supported by the Russian Foundation for Basic Research (Nos 04-01-00458, 05-04-49068, 05-07-90274), the Siberian Branch of the Russian Academy of Sciences (Interdisciplinary integration project No. 24). The authors are grateful to Vitaly Likhoshvai for valuable discussions.

REFERENCES

Demidenko G.V., Likhoshvai V.A. (2005) On differential equations with retarded argument. *Sib. Mat. Zh*., **46**, 538–552; English transl. in *Sib. Math. J.*, **46**, 417–430.

Likhoshvai V.A., Fadeev S.I., Demidenko G.V., Matushkin Yu.G. (2004) Modelling of multi-stage synthesis without branching by an equation with delay. *Sib. Zh. Ind. Mat*., **7**, 73–94. (In Russ.).

MATRIX PROCESS MODELLING: ON PROPERTIES OF SOLUTIONS OF ONE DELAY DIFFERENTIAL EQUATIONS

***Demidenko G.V.*[*1], *Khropova Yu.E.*[2]**
[1] Sobolev Institute of Mathematics, SB RAS, Novosibirsk, 630090, Russia; [2] Novosibirsk State University, Novosibirsk, 630090, Russia
[*] Corresponding author: e-mail: demidenk@math.nsc.ru

Key words: modeling of substance synthesis process, delay differential equations, initial problem, Haar functions

SUMMARY

Motivation: Many biological processes are modeled by systems of ordinary differential equations; moreover, orders of the systems may be very high. For example, synthesis of RNAs and proteins in gene networks involves hundreds and even thousands of intermediate stages. In the case of a sufficiently large number of intermediate stages, for a system modeling multi-stage substance synthesis without branching (Likhoshvai *et al.*, 2004), we showed that this system can be replaced by one delay differential equation. However, we considered the zero initial conditions for the system. Obviously, initial conditions can be arbitrary in real problems.

Results: In the present paper we obtain a generalization of the limit theorem proved in (Likhoshvai *et al.*, 2004) for the zero initial conditions in the case of arbitrary initial conditions.

INTRODUCTION

Modeling gene networks, it is necessary to take into consideration synthesis of tens and hundreds of thousands of intermediate stages of substances (DNA, RNA, proteins). Therefore, studying corresponding models, a researcher confronts with the *high dimensionality problem*.

In the paper (Likhoshvai *et al.*, 2004) the high dimensionality problem was studied for a system of ordinary differential equations modeling substance synthesis without branching

$$\begin{aligned} &\frac{dx_1}{dt} = g(x_n) - \frac{n-1}{\tau} x_1, \\ &\frac{dx_i}{dt} = \frac{n-1}{\tau}(x_{i-1} - x_i), \qquad i = 2, ..., n-1, \\ &\frac{dx_n}{dt} = \frac{n-1}{\tau} x_{n-1} - \theta x_n, \qquad \tau > 0, \quad \theta \geq 0. \end{aligned} \tag{1}$$

In the case $n >> 1$, in the paper (Likhoshvai *et al.*, 2004) a new method for finding an approximate solution of the Cauchy problem with the zero initial conditions $x\big|_{t=0} = 0$

was proposed. The main idea of this method is to enlarge unrestrictedly the system (1) and to study the limit of the sequence consisting of the last components of solutions of the Cauchy problem. As was proved (Likhoshvai *et al.*, 2004), for $g(z) \in C^1(R)$ we have a uniform convergence $x_n(t) \to y(t)$, $n \to \infty$, $t \in [0,T]$, where $y(t)$ is a solution of the following delay differential equation

$$\frac{d}{dt} y(t) = -\theta y(t) + g(y(t-\tau)), \quad t > \tau; \tag{2}$$

moreover,

$$y(t) = 0, \quad t \in [0,\tau]. \tag{3}$$

Hence, to find approximately the last component $x_n(t)$ for $n >> 1$ it is enough to solve the initial problem (2), (3). Then we obtain $x_n(t) \approx y(t)$.

In the present paper we continue to study connections between solutions of the system (1) for $n >> 1$ and solutions of the delay differential equation (2) when initial conditions are arbitrary (Likhoshvai *et al.*, 2004; Demidenko *et al.*, 2006).

RESULTS

Consider the Cauchy problem for the system (1) with initial conditions

$$x|_{t=0} = x_0. \tag{4}$$

Suppose that the function $g(z) \in C(R)$ is bounded and satisfies the Lipschitz condition. Then the problem (1), (4) has a unique solution $x(t) = (x_1(t), ..., x_n(t))^T$ on any interval $[0,T]$. Let us increase unrestrictedly the number of equations in (1) and consider only the last components $x_n(t)$ of solutions of the Cauchy problems. Then we obtain a sequence $\{x_n(t)\}$ on the interval $[0,T]$.

Theorem 1. *Let the initial conditions in the Cauchy problem (1), (4) have the form* $x_0 = (a_1, ..., a_k, 0, ..., 0)^T$, *where* k *does not depend on* n. *Then:*

a) for any $T > \tau$ *the convergence holds*

$$\left\| x_n(t) - y(t), L_p(0,T) \right\| \to 0, \quad p \geq 1, \quad n \to \infty; \tag{5}$$

b) the limit function $y(t) \in W_p^1(\tau, T)$ *is a weak solution of the initial problem for the delay differential equation*

$$\begin{aligned} &\frac{d}{dt} y(t) = -\theta y(t) + g(y(t-\tau)), \qquad t > \tau, \\ &y(t) = 0, \qquad t \in [0,\tau), \qquad y(\tau + 0) = \sum_{i=1}^{k} a_i; \end{aligned} \tag{6}$$

c) the function $y(t)$ *is a classic solution of the equation (2) for* $t > 2\tau$.

Note that the condition indicated in Theorem 1 on the initial vector x_0 is essential. If we will consider initial vectors with finite number of nonzero components a_k, where k depends on *n*, then we can establish the convergence (5) to a weak solution of the delay differential equation (2). However, initial conditions will differ from the initial conditions in (6). We give a few examples below.

For simplicity, we assume that $\theta = 0$, $\tau = 1$. In the next theorems we consider a few examples of initial conditions in the Cauchy problem (1), (4). These examples show that, as the number of equations increases unrestrictedly, the last components $x_n(t)$ tend to weak solutions of the equation (2) with initial conditions of the form

$$y(t) = \varphi_{m,k}(t), \quad 0 \le t \le 1, \quad m \in N, \quad k = 1, 2, ..., 2^m. \tag{7}$$

where

$$\varphi_{m,k}(t) = \begin{cases} 2^{m/2}, & t \in [(k-1)/2^m, (k-1/2)/2^m), \\ -2^{m/2}, & t \in [(k-1/2)/2^m, k/2^m), \\ 0, & t \notin [(k-1)/2^m, k/2^m). \end{cases} \tag{8}$$

Note that the system of functions $\{\varphi_{m,k}(t)\}$ forms the orthonormal Haar basis in the space $L_2(0,1)$ (see, for instance, (Triebel, 1983)).

Theorem 2. *Let* $n = 4l$. *Assume that the initial vector in the Cauchy problem (1), (4) has components*

$a_{n/2} = \sqrt{2}$, $a_{3n/4} = -2\sqrt{2}$, $a_n = \sqrt{2}$, *the rest of components* $a_k = 0$,

or

$a_{n/4} = -2\sqrt{2}$, $a_{n/2} = \sqrt{2}$, *the rest of components* $a_k = 0$.

Then, for any $T > \tau$, *the convergence (5) holds and the limit function* $y(t)$ *is a weak solution of the equation (2) with the initial conditions* $\varphi_{1,1}(t)$ *or* $\varphi_{1,2}(t)$, *respectively.*

Theorem 3. *Let* $n = 8l$. *Assume that the initial vector in the Cauchy problem (1), (4) has components*

$a_{3n/4} = 2$, $a_{7n/8} = -4$, $a_n = 2$, *the rest of components* $a_k = 0$,

or

$a_{n/2} = 2$, $a_{5n/8} = -4$, $a_{3n/4} = 2$, *the rest of components* $a_k = 0$,

or

$a_{n/4} = 2$, $a_{3n/8} = -4$, $a_{n/2} = 2$, *the rest of components* $a_k = 0$,

or

$a_{n/8} = -4$, $a_{n/4} = 2$, *the rest of components* $a_k = 0$.

Then, for any $T > \tau$, *the convergence (5) holds and the limit function* $y(t)$ *is a weak solution of the equation (2) with the initial conditions* $\varphi_{2,1}(t)$ *or* $\varphi_{2,2}(t)$ *or* $\varphi_{2,3}(t)$ *or* $\varphi_{2,4}(t)$, *respectively.*

Analogous assertions can be formulated for the Cauchy problem for $n = 2^{m+1}l$, $\theta = 0$, $\tau = 1$. Thus, one can establish that $x_n(t) \approx y(t)$ for $n >> 1$, and $y(t)$ is a weak solution of the equation (2) with initial conditions of the form (7).

DISCUSSION

Using analogs of the formulated theorems, we can point out vectors x_0 of initial conditions in (4) under which the last components of solutions of the Cauchy problems of the form (1), (4) for $n >> 1$ approximate solutions of delay differential equations of the form (2) with initial conditions of the form

$y(t) = \varphi(t)$, $t \in [0,1]$, $\varphi(t) \in C[0,1]$. (9)

To this end, it is sufficient to write an expansion of the function $\varphi(t)$ in a series in the Haar functions (8) in $L_2(0,1)$

$$\varphi(t) = \sum_{m=1}^{\infty} \sum_{k=1}^{2^m} c_{m,k} \varphi_{m,k}(t), \tag{10}$$

where

$$c_{m,k} = \int_0^1 \varphi_{m.k}(s)\varphi(s)ds. \tag{11}$$

For any $\varepsilon > 0$ we can find a number m_ε such that

$$\left\| \varphi(t) - \sum_{m=1}^{m_\varepsilon} \sum_{k=1}^{2^m} c_{m,k} \varphi_{m,k}(t), L_2(0,1) \right\| \le \varepsilon / 2.$$

Consider the Cauchy problem (1), (4) for $n = 2^{m_\varepsilon + 1}l$. Using analogs of Theorem 1–3, taking into account (10), (11), we choose a relevant initial vector x_0. Then, in the same way as in (Demidenko *et al.*, 2006), for any $\varepsilon > 0$ there exists n_ε such that, for any $n > n_\varepsilon$, the last component of a solution of the Cauchy problem (1), (4) satisfies the inequality

$|x_n(t) - y(t)| \le \varepsilon$, $t \in [\tau, T]$,

where $y(t)$ is a solution of the initial problem (2), (9).

Our result makes it possible to solve the high dimensionality problem for the system of ordinary differential equations (1) modeling substance synthesis without branching in the case of arbitrary initial conditions. We plan to extend the result to other systems modeling gene networks.

ACKNOWLEDGEMENTS

The work was supported by the Russian Foundation for Basic Research (Nos 04-01-00458) and the Siberian Branch of the Russian Academy of Sciences (Interdisciplinary integration project No. 24). The authors are grateful to V.A. Likhoshvai for helpful discussions.

REFERENCES

Demidenko G.V., Likhoshvai V.A., Kotova T.V., Khropova Yu.E. (2006) On one class of systems of differential equations and on retarded equations. *Sib. Mat. Zh.*, **47**, 58–68; English transl. in *Sib. Math. J.*, **47**, 45–54.

Likhoshvai V.A., Fadeev S.I., Demidenko G.V., Matushkin Yu.G. (2004) Modelling of multi-stage synthesis without branching by an equation with delay. *Sib. Zh. Ind. Mat.*, **7**, 73–94. (In Russ.).

Triebel H. (1983) Theory of Function Spaces. Monographs in Mathematics, Vol. 78. Birkhäuser Verlag, Basel-Boston-Stuttgart.

ASYMPTOTIC PROPERTIES OF SOLUTIONS OF DIFFERENTIAL-DIFFERENCE EQUATIONS WITH PERIODIC COEFFICIENTS IN LINEAR TERMS

Demidenko G.V.*[*], *Matveeva I.I.
Sobolev Institute of Mathematics, SB RAS, Novosibirsk, 630090, Russia
[*] Corresponding author: e-mail: demidenk@math.nsc.ru

Key words: gene networks, blood cell production, mathematical models, delay differential equations, asymptotic stability, attraction domain

SUMMARY

Motivation: Many biological processes are modeled by delay differential equations. Therefore, for more deep insight into these processes, it is necessary to develop effective theoretical and numerical methods to study qualitative properties of solutions of such equations.

Results: In the present paper we give results concerning asymptotic stability of solutions of quasilinear systems of delay differential equations with periodic coefficients in linear terms. From these results we obtain conditions for asymptotic stability of stationary solutions of equations modeling substance synthesis and the blood cell production.

INTRODUCTION

Various biological processes are modeled by delay differential equations. For example, such equations arise when modeling of gene networks, population dynamics, the blood cell production, and etc. Therefore development of methods for study of qualitative properties of solutions of such equations has theoretical as well as practical importance.

The present paper is devoted to delay differential equations with periodic coefficients in linear terms

$$\frac{d}{dt}y(t) = A(t)y(t) + B(t)y(t-\tau) + F(y(t), y(t-\tau)), \quad t > \tau > 0, \tag{1}$$

where $A(t)$, $B(t)$ are $n \times n$ matrices with continuous periodic entries, i.e.

$$A(t+T) = A(t), \quad B(t+T) = B(t), \quad T > \tau,$$

$F(u,v)$ is a real-valued vector-function satisfying the Lipschitz condition and

$$\|F(u,v)\| \le q_1 \|u\|^{1+\omega_1} + q_2 \|v\|^{1+\omega_2}, \quad q_1, q_2, \omega_1, \omega_2 \ge 0.$$

Our goal is to study asymptotic stability of the zero solution of the system (1) and obtain estimates characterizing decay rate as $t \to \infty$. These results generalize the authors' results from the paper (Demidenko, Matveeva, 2005).

In (Demidenko, Matveeva, 2005) a modification of the Lyapunov-Krasovskii functional was proposed. Using the modification, we obtained estimates of exponential decrease at infinity for solutions of linear and quasilinear systems of delay differential equations of the form (1) with constant coefficients. Attraction domains of the zero solution were established too.

Using the authors' results from (Demidenko, Matveeva, 2001), in the present paper we obtain analogous results for linear and quasilinear systems of delay differential equations of the form (1) in the case of periodic coefficients.

RESULTS

At first, we consider the following linear system of delay differential equations with *T*-periodic coefficients

$$\frac{d}{dt}y(t) = A(t)y(t) + B(t)y(t-\tau), \quad t > \tau. \tag{2}$$

In the next theorem we give sufficient conditions of asymptotic stability of the zero solution of the system and establish estimates for solutions of (2).

Theorem 1. *Suppose that there exist matrices*

$H(t) = H^*(t) \in C^1[0,T]$ *and* $K(s) = K^*(s) \in C^1[0,\tau]$

such that

$$H(0) = H(T) > 0, \quad K(s) > 0, \ \frac{d}{ds}K(s) < 0, \quad s \in [0,\tau],$$

and the matrix

$$C(t) = -\begin{pmatrix} \dfrac{d}{dt}H(t) + H(t)A(t) + A^*(t)H(t) + K(0) & H(t)B(t) \\ B^*(t)H(t) & -K(\tau) \end{pmatrix}$$

is positive definite on the interval $[0,T]$. *Denote by* $c_1(t) > 0$ *the minimal eigenvalue of the matrix* $C(t)$ *and by* $k > 0$ *the maximal number such that*

$$\frac{d}{ds}K(s) + kK(s) \le 0, \qquad s \in [0,\tau].$$

Then the zero solution of the system (2) is asymptotically stable; moreover, for a solution of (2) with an initial function $\varphi(t) \in C[0,\tau]$, *the following inequality holds*

$$\|y(t)\|^2 \le h_1^{-1}(t)\exp\left(-\int_\tau^t \frac{\varepsilon(\xi)}{\|H(\xi)\|}d\xi\right)\left[\langle H(\tau)\varphi(\tau),\varphi(\tau)\rangle + \int_0^\tau \langle K(\tau-s)\varphi(s),\varphi(s)\rangle ds\right],$$

$t > \tau$,

where $h_1(t) > 0$ *is the minimal eigenvalue of the matrix* $H(t)$,

$$\varepsilon(t) = \min\{c_1(t), kH(t)\}. \tag{3}$$

We now consider the quasilinear system (1). For simplicity, we formulate our result in the case $q_2 = 0$.

Theorem 2. *Let the conditions of Theorem 1 be satisfied and*

$$r^{\omega_1/2} = \left(1 - \exp\left(-\frac{\omega_1}{2}\int_0^T \frac{\varepsilon(s+\tau)}{\|H(s+\tau)\|}\,ds\right)\right)\times$$

$$\times\left(q_1\omega_1\int_0^T \frac{\|H(\xi+\tau)\|}{h_1^{1+\omega_1/2}(\xi+\tau)}\exp\left(-\frac{\omega_1}{2}\int_0^\xi \frac{\varepsilon(s+\tau)}{\|H(s+\tau)\|}\,ds\right)d\xi\right)^{-1},$$

where $\varepsilon(t) > 0$ *is given by (3),* $h_1 > 0$ *is the minimal eigenvalue of the matrix* $H(t)$. *Then the zero solution of the system (1) is asymptotically stable, and the set of real-valued functions*

$$\Sigma = \left\{\varphi(t) \in C[0,\tau] : \left\langle H(\tau)\varphi(\tau), \varphi(\tau)\right\rangle + \int_0^\tau \left\langle K(\tau - s)\varphi(s), \varphi(s)\right\rangle ds < r\right\}$$

is an attraction domain of the zero solution. Moreover, for a solution of the system (1) with an initial function $\varphi(t) \in \Sigma$, *the following estimate holds*

$$\|y(t)\|^2 \le h_1^{-1}(t)\exp\left(-\int_\tau^t \frac{\varepsilon(\xi)}{\|H(\xi)\|}\,d\xi\right)\left[\left\langle H(\tau)\varphi(\tau), \varphi(\tau)\right\rangle + \int_0^\tau \left\langle K(\tau - s)\varphi(s), \varphi(s)\right\rangle ds\right]\times$$

$$\times\left(1 - r^{-\omega_1/2}\left[\left\langle H(\tau)\varphi(\tau), \varphi(\tau)\right\rangle + \int_0^\tau \left\langle K(\tau - s)\varphi(s), \varphi(s)\right\rangle ds\right]^{\omega_1/2}\right)^{-2/\omega_1}, \quad t > \tau.$$

DISCUSSION

It is well-known that, for systems of linear delay differential equations with constant coefficients ($A(t) = A$, $B(t) = B$, $F(u,v) = 0$), there is an analog of the spectral theorem on asymptotic stability. Namely, if all roots of the quasipolynomial $\det(A + e^{-\lambda\tau}B - \lambda I) = 0$ belong to the left half-plane $C_- = \{\lambda \in C : \operatorname{Re}\lambda < 0\}$, then the zero solution is asymptotically stable. However, it is very hard to verify the condition. In the case of periodic coefficients an analogous spectral problem becomes practically unsolvable.

The results obtained in this paper make it possible to study qualitative properties of solutions of delay differential equations without finding roots of quasipolynomials. It provides ample opportunities to conduct computational investigations and obtain qualitative characteristics for biological processes.

We illustrate one of our results by the delay differential equation encountered in various biological problems (for modeling of gene networks (Likhoshvai *et al*., 2004), for describing the blood cell production (Kuang, 1993)).

$$\frac{d}{dt}y(t) = -\theta y(t) + \frac{\alpha\beta^\gamma}{\beta^\gamma + y^\gamma(t-\tau)}, \quad t > \tau, \tag{4}$$

where α, β, $\theta > 0$, $\gamma > 0$ is integer. Consider a positive stationary solution y_c of the equation (4). It is unique and defined by the equality

$$\theta y_c = \frac{\alpha\beta^\gamma}{\beta^\gamma + y_c^\gamma}.$$

From our results it follows that the stationary solution is asymptotically stable for $\theta > \frac{\alpha\gamma}{\beta}$.

Development of new methods for analysis of models describing dynamics of biological processes is one of actual problems of systemic biology. Many biological systems are modeled by means of delay differential equations. Examples of such models can be found in many books (see, for example, Murray, 1977; Marchuk, 1983; Kuang, 1993). In future, we plan to extend our results to various models.

ACKNOWLEDGEMENTS

The work was supported by the Russian Foundation for Basic Research (No. 04-01-00458), and the Siberian Branch of the Russian Academy of Sciences (Interdisciplinary integration project No. 24).

REFERENCES

Demidenko G.V., Matveeva I.I. (2001) On stability of solutions of linear systems with periodic coefficients. *Sib. Mat. Zh.*, **42**, 332–348; English transl. in *Sib. Math. J.*, **42**, 282–296.

Demidenko G.V., Matveeva I.I. (2005) Asymptotic properties of solutions of delay differential equations. *Vestnik NGU. Ser. matematika, mekhanika, informatika*, **5**, 20–28.

Kuang Y. (1993) Delay Differential Equations with Applications in Population Dynamics. *Mathematics in Science and Engineering*. Vol. 191. Academic Press, Boston.

Likhoshvai V.A., Demidenko G.V., Fadeev S.I., Matushkin Yu.G., Kolchanov N.A. (2004) Mathematical simulation of regulatory circuits of gene networks. *Zh. Vychisl. Mat. Mat. Fiz.*, **44**, 2276–2295; (In Russ.). English transl. in *Comput. Math. Math. Phys*., **44**, 2166–2183.

Marchuk G.I. (1983) *Mathematical Models in Immunology*. Optimization Software, Inc., Publications Division, New York.

Murray J.D. (1977) *Lectures on Nonlinear-Differential-Equation Models in Biology*. Clarendon Press, Oxford.

PROGRAM PACKAGE HGNET FOR COMPUTATIONAL STUDIES OF HYPOTHETICAL GENE NETWORKS

***Fadeev S.I.*[*1], *Korolev V.K.*[1]**
[1] Sobolev Institute of Mathematics, SB RAS, Novosibirsk, 630090, Russia
[*] Corresponding author: e-mail: fadeev@math.nsc.ru

Key words: genetic systems, modeling, differential autonomous systems, regulator circuit of gene networks, hypothetical gene networks

SUMMARY

Motivation: Intense accumulation of data on the structure, function, and dynamic behavior of gene networks put the investigation of structural and functional regularities of gene networks in the forefront. For this purpose, natural and artificial gene networks should be studied both theoretically and experimentally. In particular, better understanding of the relationships between structural and dynamic features of gene networks demands their study *in silico* (Likhoshvai *et al.*, 2003).

Results: In this paper we present the HGNET program package, designed for studies of mathematical models describing four classes of hypothetical gene networks introduced by Likhoshvai *et al.* (2003).

INTRODUCTION

Intense accumulation of data on the structure, function, and dynamic behavior of gene networks put the investigation of structural and functional regularities of gene networks in the forefront. These regularities endow gene networks with ability to operate in different steady-state and/or oscillatory regimes. Problems formulated in (Likhoshvai *et al*., 2003; Likhoshvai, Fadeev, 2003; Likhoshvai *et al*., 2004) are reflected in the HGNET package. Functional regularities of gene network regulator circuits are studied in hypothetical constructions called hypothetical gene networks (HGNs), which are modeled by special classes of autonomous systems and systems with retarded arguments. The package is intended for investigation of models belonging to this class. It uses both original algorithms, whose effectiveness is determined by properties of models under consideration, and general algorithms reported in (Fadeev *et al.*, 1998). The package has been developed in Visual Basic 6. By now, the package is supplemented with a unit that allows numerical investigation of equation sets for HGNs with regulatory connections giving rise to chaotic oscillations.

FUNCTION OF THE PACKAGE

The HGNET program package has been developed for investigation of HGN models with regard to their dynamics, limiting states (steady-state conditions and limit cycles), and limiting state stability.

Units of the HGNET package perform the following numerical studies: (1) numerical integration in a problem with initial data for autonomous equation sets whose right sides model regulatory relations in HGNs of classes 1–4; (2) numerical integration in a problem with initial data for corresponding equations with retarded arguments; (3) calculation of steady-state solutions and determination of their multiplicity by the method of parameter continuation with simultaneous determination of their numerical stability by the Godunov-Bulgakov method; (4) analysis of stable and unstable oscillations; (5) integration of autonomous systems modeling template-directed processes with numerous intermediate stages.

The *graphic interface* of HGNET supports interaction between a use and program units. The user can obtain graphic presentation of the results of autonomous system integration in physical and phase planes, Poincare maps, projections of the three-dimensional phase space and the Poincare hyperplane. Also there are options for construction of stereoscopic images and rotation of three-dimensional objects.

RESULTS

Consider several examples of HGNET application.

Example 1. One of the topical problems of the gene network theory is the understanding of the role of gene network structure in their dynamic properties. For computational studies in this field, HGNET contains units allowing study of models of the four HGN classes determined in (Likhoshvai *et al.*, 2003). For example, a user can perform numerical analysis of the system of n equations describing an HGN with regulatory connections of class 1:

$$dx_i / dt = \alpha /(1+\beta \sum_{j=1}^{n} S_{i,j} x_j^{\gamma}) - x_i, \quad i = 1, 2, ..., n,$$

where $S_{i,j}$ are elements of the connection matrix, and $\alpha > 0$, $\beta > 0$, $\gamma > 1$ are parameters. For symmetrical HGNs (M(n,k) model), the connection matrix is defined by specifying the integer parameter k, $1 < k \le n$ (Likhoshvai *et al.*, 2003).

The three-dimensional stereoscopic image in Fig. 1 shows the achievement of the steady-state limiting cycle by an M(3,2) model, represented by equations:

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

the values of parameters being $\alpha = 5$, $\beta = 1$, $\gamma = 4$. The variation range of parameter a allowing self-excited oscillations is determined by analysis of the stability of the steady-state solution depending on a.

This model is the least complex cyclic system of three genetic elements. It describes a commonly known genetic repressilator, embodied by gene engineering methods (Elowitz, Leibler, 2000). Experiments showed that such genetic constructs could oscillate. So can model (1). Of special interest is the parabolic approach of the 3D curve to the limiting cycle of subtriangular shape.

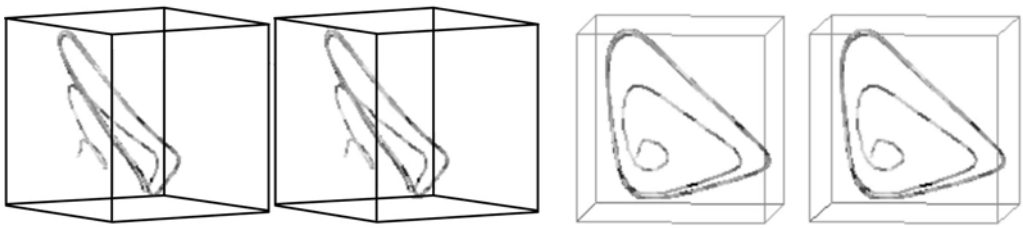


Figure 1. Stereoscopic presentation of the achievement of the steady-state limiting cycle by an M(3,2) model at two angles in the phase plane.

Example 2. Consider a model for synthesis of linear molecules: protein, RNA, or DNA (Fadeev *et al.*, 2005):

$$
\begin{aligned}
&a = (n-1)/\tau_1,\ b = (n-1)/\tau_1 \quad S = a+b+\omega,\ f(y_n) = 1/(1+\beta y_n^{\gamma}).\\
&dy_1/dt = -(a+\omega)y_1 + b\,y_2 + \alpha\, f(y_n),\\
&dy_i/dt = a\,y_{i-1} - S\,y_i + b\,y_{i+1},\ i = 1,2,...,n-2,\\
&dy_{n-1}/dt = a\,y_{n-2} - S\,y_{n-1},\ dy_n = a y_{n-1} - \theta\, y_n,
\end{aligned}
\tag{2}
$$

Here y_n is the concentration of the target product; τ_1, τ_2 are summarized durations of stages between the 1st and nth states; α is the constant of the elongation rate of the molecule in the direct process; β, constant of the rate of molecule shortening in the reverse process; ω, rate of spontaneous termination of molecule elongation; and θ, degradation rate of the target product.

The methodological significance of this model type stems from the basic problem of relations between various levels of system organization in the course of its operation. For template processes, study of the conditions for correct transition from type (2) systems to Eq. 3

$$
t > \tau,\quad dx/dt = \alpha\, f(x(t-\tau))\, e^{-\omega\tau} - \theta\, x, \tag{3}
$$

also describing this process, presents a problem.

Such studies involve numerical integration of high-dimensionality systems. The HGNET packages implements a semiimplicit method for integration of systems (2) and Eq. (3), allowing efficient numerical studies and comparison of solutions.

Fig. 2 illustrates self-excited oscillations in system (2) with parameters $\alpha = 5,\ \beta = 1, \gamma = 5, \tau_1 = 2, \tau_2 = 5,\ \theta = 2,\ \omega = 0.1$. The number of intermediate stages n = 1000. The amplitude of intermediate component oscillations is close to zero. The curve for component y_{1000} is virtually the same as the $x(t)$ curve, which is a solution of Eq. (3), where the value of the retarded argument is calculated as $\tau = \tau_1\tau_2/(\tau_2 - \tau_1),\ \tau_2 > \tau_1$. This is in agreement with the statement of the limiting theorem of uniform approach of $y_n(t)$ to $x(t)$ at $n \to \infty$ (Fadeev *et al.*, 2006).

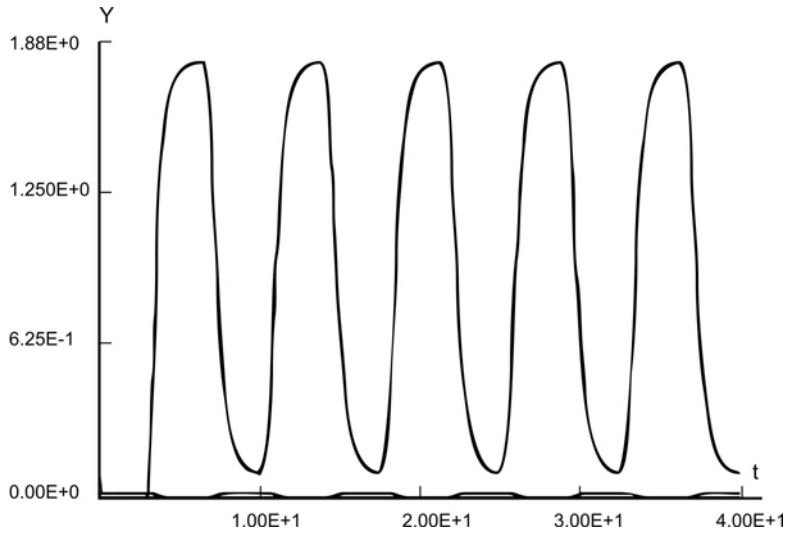


Figure 2. Self-excited oscillations in the model of synthesis with a large number of intermediate stages (n = 1000). The components of intermediate stages are close to zero.

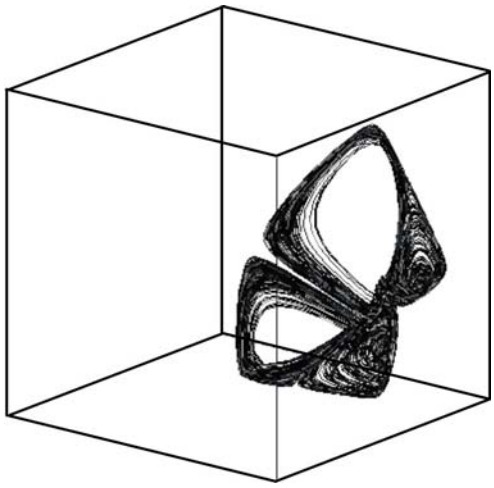


Figure 3. Phase-space representation of chaotic oscillations obtained by numerical integration of system (4) at m = 3, α = 8.5, and γ = 4.

The model is represented by the following autonomous set of equations:

$$dx_1/dt = P^{(m)}(u_1) - x_1\ ,\ dx_2/dt = P^{(m)}(u_2) - x_2\ ,dx_3/dt = P^{(m)}(u_3) - x_3\ ,(4)$$

where $u_1 = x_2 + x_3$, $u_2 = x_3 + x_1$, $u_3 = x_1 + x_2$,

$$P^{(m)}(z) = \alpha\frac{P^{(m-1)}(z)}{1+[P^{(m-1)}(z)]^{\gamma}},\ m = 2,3,.....,\ P^{(1)}(z) = \frac{\alpha}{1+z^{\gamma}}.\ m = 1,2,....$$

The properties of this system are considered in more detail in (Likhoshvai *et al.*, 2006).

CONCLUSIONS

The HGNET software reported here is constructed for supporting *in silico* experiments in the theory of gene networks. According to its intent, we develop it as a store for solved problems of this theory and a tool for analyzing special classes of mathematical models describing various types of artificial gene constructs. The package can be applied to practical construction of gene networks with preset properties. The authors are grateful to Vitaly Likhoshvai for valuable discussions.

ACKNOWLEDGEMENTS

The authors are grateful to Nikolay Kolchanov and Vitaly Likhoshvai for valuable discussion and to Victor Gulevich for translating the manuscript from Russian into English. This work was supported in part by the Russian Foundation for Basic Research (No. 06-04-49556).

REFERENCES

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

Likhoshvai V.A., Fadeev S.I. (2003) About hypothetical gene networks. *Sib. J. of Industrial Mathematics*, **4**, 134–153.

Likhoshvai V.A., Fadeev S.I., Demidenko G.V., Matushkin Yu.G. (2004) Modeling of multi-stage synthesis without bifurcation by the equation with retarded argument. *Sib. J. of Industrial Mathematics*, **7**, 73–94.

Fadeev S.I., Pokrovskaya S.A., Berezin A.Yu., Gainova I.A. (1998) *Program Package STEP for Computational Investigation Nonlinear Equations Sets and Autonomous Systems in General Form*. Novosibirsk, NSU, 188 p.

Fadeev S.I., Shtokalo D.N. Likhoshvai V.A. (2006) Matrix process modelling: study of a model of synthesis of linear biomolecules with regard to reversibility of processes. This issue.

Elowitz M.B., Leibler S. (2000) A synthetic oscillatory network of transcriptional regulators. *Nature*, **403**(6767), 335–338.

Fadeev S.I., Likhoshvai V.A., Shtokalo D.N. (2005) Study of a model for synthesis of linear biomolecules with regard to reversible processes. *Sib. J. of Industrial Mathematics*, **8**, 149–162.

Likhoshvai V.A., Rudneva D.S., Fadeev S.I. (2006) Oscillations of chaotic type in symmetric gene networks of small dimension. *This issue.*

GENE NETWORKS BEHAVIOR IN A SERIES OF SUCCESIVE CELL DIVISIONS

Galimzyanov A.V.
Department of Physicochemical Biology and Epigenetics, Ufa Research Center, RAS, Ufa, Russia,
e-mail: galim@anrb.ru

Key words: gene networks, cell division, modeling

SUMMARY

Motivation: When modeling the dynamics of the gene networks in developing cell ensembles, one should account for the processes of intercellular 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: A computer model is constructed for the process of gene product distribution among daughter cells during the maternal cell division. An algorithm is proposed for traversing the tree of cell divisions with unique gene networks that makes it possible to obtain gene expression profiles in the cells of intermediate and final generations. Peculiar features are analyzed in the functioning of a gene network, which includes three cyclic digene systems with negative feedback, in successive cell generations.

Availability: The MAGeneNT software is available on request from the author.

INTRODUCTION

In successive cell generations, the transmission of extra-genome regulatory molecules, e.g., proteins, proceeds from ancestor to descendant. The gene-expression profile in the maternal cell before division determinates the initial data set of gene networks (GN) in the daughter cells. In this case the gene subnetwork dynamics can be either changed or preserved. In this connection in order to better appreciate the mechanisms controlling the development of living systems one should be able to trace the GN dynamics within cell lines.

METHOD AND ALGORITHM

Distribution model of substance's molecules among daughter cells

The division of cell $c_{i,d}$ (i is the cell number in the d-th generation) gives rise to two daughter cells $c_{i_1,d+1}$ and $c_{i_2,d+1}$. Molecules of the products of the j-th gene enter cell $c_{i_1,d+1}$ with probabilities $p_{m,j}$ (for each RNA molecule of the j-th gene) and $p_{r,j}$ (for each protein molecule of the j-th gene) or they enter cell $c_{i_2,d+1}$ with probabilities $q_{m,j}$ and $q_{r,j}$. Each molecule has two possible outcomes (entry into one of the two daughter cells), the probabilities of outcomes therewith remain constant ($p = q = ½$). Thus, for each cell $c_{i,d}$ ($i = \overline{1,2^d}$, $d = \overline{1,D}$, where D is the number of generations under observation) and some substance $\Re$ we have the Bernoulli distribution of n trials, where n is the amount of the molecules of substance $\Re$ in a cell.

The number of successes S_n in n trials is a random value with the binomial distribution function. In our case it is interesting to compute not the probability of getting exactly k successes, but to find, with high probability (0.99), the range (α, β), in which value S_n lies. According to the de Moivre-Laplace limit theorem, for fixed x_1 and x_2 at $n\to\infty$ we have: $P\{np + x_1\sqrt{npq} \le S_n \le np + x_2\sqrt{npq}\} \to \Phi(x_2) - \Phi(x_1)$, where $x_1 = (\alpha - np)\,h$, $x_2 = (\beta - np)\,h$, $h = 1/\sqrt{npq}$, $\Phi(x) = (1/\sqrt{2\pi}) \int_{-\infty}^{x} e^{-y^2/2}\partial y$ (the function is tabulated).

In the problem under consideration $x_1<0$, $x_2>0$, $x_2 = |x_1|$, therefore $\Phi(x_2) - \Phi(x_1) = 2\Phi(x_2) - 1$; at P = 0.99: $\Phi(x_2) = 0.995$, $x_2 = 2.61$. Hence, the range (α, β) is defined, in which we randomly select the value of variation $\Delta_{1/2}$ from the half-value concentration of substance $\Re$ in a cell before division (at the end of life cycle).

Algorithm "Greence" for walking through a tree of cell divisions

The algorithm is based upon a ripple-through tracing cell lines from the original parent cell to all cells of the final generation (Fig. 1). A V-element is introduced into consideration to walk through a tree of cell divisions, with encountered nodes being discovered one by one. To discover node $\upsilon_{i,d}$ (i is the number of a cell in the d-th generation) means to calculate the GN dynamics in two daughter cells $c_{i_1,d+1}$ and $c_{i_2,d+1}$ originating from parent cell $c_{i,d}$ that corresponds to a given node. According to concentration levels of regulatory substances at the end of life cycles of cells $c_{i_1,d+1}$ and $c_{i_2,d+1}$, GN initial data is produced in their derivative cells, i.e., $c_{i_{1_1},d+2}$, $c_{i_{1_2},d+2}$ and $c_{i_{2_1},d+2}$, $c_{i_{2_2},d+2}$, respectively. The nodes ($\upsilon_{i_1,d+1}$ and $\upsilon_{i_2,d+1}$) that represent such daughter cells ($c_{i_1,d+1}$ and $c_{i_2,d+1}$) are said to be actualized. Next only one of the daughter cells (either $c_{i_1,d+1}$ or $c_{i_2,d+1}$) is taken as a parent, and its representing node is discovered. The other node gains the status of a delayed alternative and simply waits for its turn. It will be discovered at a later time, and then the walk will begin through the branches running from this already discovered node to the nodes of the upper level. On reaching the final layer (cell generation), the V-element comes back along the same fragment of the ripple-through path already finished to the nearest delayed node (it is actualized) and begins to complete the delayed alternative ripple-through path.

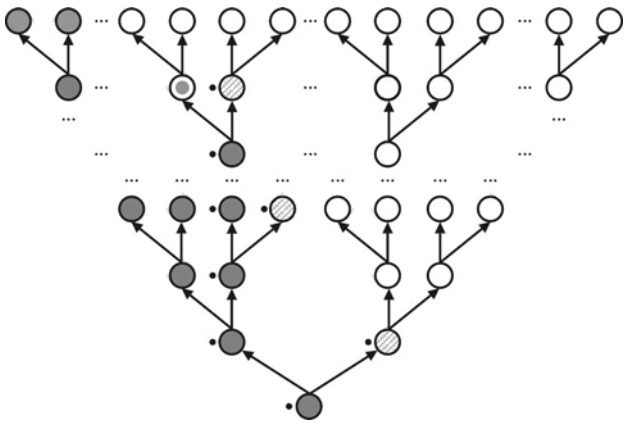


Figure 1. Walking through the graph of cell divisions with the use of recursion. Designations: circle with dark inset refers to the *V*-element, white circle to undiscovered node, dark circle to discovered node; shaded circle to delayed node; marks (•) near the circles show a memorized route to delayed nodes; $d = \overline{1, D}$ is the number of generation.

The *V*-element is an argument and output of the recursive function that determines the direction of its next step and the type of operations to be performed on the encountered node (whether it should be actualized, discovered, memorized as a point of the route or cut off). The use of recursion makes it unnecessary to store the information on each cell of the generation preceding a new one. The *V*-element "brings" initial data for a new step and also leaves them in the delayed nodes of an alternative route (not more than the number of generations); subtrees with all nodes discovered and without delayed nodes are cut off. The approach described above enables one to calculate the dynamics of intracellular GN in successive cell generations up to 30 in number comparable with Hayflick's limit (~50) for a number of eukaryotic cell divisions (Hayflick, Moorhead, 1961).

IMPLEMENTATION AND RESULTS

The model and algorithm are software realized in the concept of object-oriented programming and, as computing modulus, included in the software package MAGeneNT (Galimzyanov, 2005). The MAGeneNT system is intended for analyzing the dynamics of control gene networks based on the method of generalized threshold models (GTM) (Tchuraev, 1991) and makes it possible to describe the processes of positional information distribution, functioning of intracellular GN with the account for the values of regulatory substance concentrations, transmittance in a series of successive generations of the cells of extra-genomic regulatory molecules as well as intercellular interactions.

The approach developed by this study is applied to gene network $TCDS^{(-)}$, which includes three cyclic digene systems with negative feedback, $CDS^{(-)}$, being a genetic toggle switch (Fig. 2). The $TCDS^{(-)}$ model is constructed in terms of GTM formalism (Tchuraev, 1991). The model takes into account the following parameters: $m_j(t)$ and $r_j(t)$ are the concentrations of mRNA and proteins, respectively, expressed as the numbers of molecules per cell ($j = \overline{1,7}$); a_{1j} is the unit intensities of transcription (the promoter force) (molec./cell × min^{-1}); a_{2j} is the unit intensities of translation (molec./cell × min^{-1}); b_{1j} is the rates of mRNA degradation (molec./cell × min^{-1}); b_{2j} is the rate of protein degradation (molec./cell × min^{-1}); P_{ij}, constants describing threshold concentrations of protein product of *i*-th gene, that are necessary to inhibit the synthesis of the transcripts of *j*-th genes (these constants are determined by the affinity of regulatory substances to the protein-binding sites and by the degree of multimery); T is the cell cyclic duration (min); $K = 10$ is the number of cell generations. The parameter values for the typical $CDS^{(-)}$: $a_{11} = 2.3$, $a_{12} = 1.3$; $a_{21} = 1.5$, $a_{22} = 1.0$; $b_{11} = 0.3$, $b_{12} = 0.22$; $b_{21} = 0.013$, $b_{22} = 0.0115$, $P_{12} = 10$, $P_{21} = 9$, $T = 36$. Initial conditions: $m_1(t_0) = 7$, $m_2(t_0) = 6$, $r_1(t_0) = 300$, $r_2(t_0) = 250$. The kinetic curves for the molecular components (mRNA and proteins) of the GN were calculated individually for every cell of each generation. The initial protein and mRNA concentrations for each gene in two daughter cells were taken to be half the concentrations of these substances in the parent cell at the end of the life cycle (before division), with the account for random deviation according to the binomial law.

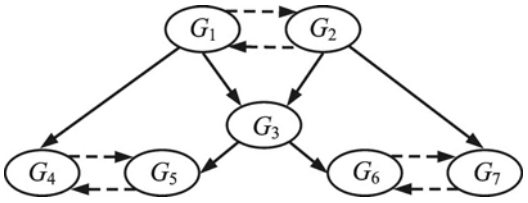


Figure 2. Scheme for the gene network $TCDS^{(-)}$. The negative interaction among genes (transcription repression) is denoted with a dotted line, and the positive interaction (transcription activation) with a solid line.

In the experiments *in silico* we obtained quantitative relations among the cells being heterogeneous by the $TCDS^{(-)}$ functional states, in a series of generations. As shown in Table 1, after a series of cell divisions cell subpopulations with four different epigenotypes occur and persist in succeeding generations.

Table 1. The relation among the cells being heterogeneous by the $TCDS^{(-)}$ functional states, in a series of generations

d	Functional states											
	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}	S_{11}	S_{12}
1	100	0	0	0	0	0	0	0	0	0	0	0
2	100	0	0	0	0	0	0	0	0	0	0	0
3	100	0	0	0	0	0	0	0	0	0	0	0
4	0	12.5	12.5	0	12.5	25	0	12.5	12.5	0	0	12.5
5	0	0	0	12.5	37.5	0	25	12.5	0	12.5	0	0
6	0	0	0	25	37.5	0	25	0	0	12.5	0	0
7	0	0	0	25	37.5	0	25	0	0	1.563	10.94	0
8	0	0	0	25	37.5	0	25	0	0	0	12.5	0
9	0	0	0	25	37.5	0	25	0	0	0	12.5	0
10	0	0	0	25	37.5	0	25	0	0	0	12.5	0

The functional states: $S_1 = (0; 0; 1; 0; 0; 0; 0)$, $S_2=(0; 1; 1; 0; 1; 0; 0)$, $S_3=(1; 0; 1; 1; 0; 0; 0)$, $S_4 = (0; 1; 1; 0; 1; 1; 0)$, $S_5 = (1; 0; 1; 1; 0; 1; 0)$, $S_6 = (0; 1; 1; 0; 0; 0; 1)$, $S_7=(0; 1; 1; 0; 1; 0; 1)$, $S_8 = (0; 1; 1; 0; 0; 1; 0)$, $S_9 = (0; 0; 1; 0; 1; 0; 1)$, $S_{10} = (1; 0; 1; 0; 1; 0; 0)$, $S_{11} = (1; 0; 1; 0; 1; 1; 0)$, $S_{12} = (1; 0; 1; 0; 0; 1; 0)$, where j-th component of S_k ($k = \overline{1,12}$) is functional state of j-th gene (1 – active, 0 – non-active) ($j = \overline{1,7}$); d – is the number of generation.

DISCUSSION

Earlier experiments *in silico* were carried out on mathematical model of $CDS^{(-)}$, being a dynamic epigene, that is a cyclic system of genes with more than one inherited functional state, or epigenotype (Tchuraev *et al.*, 2006). A new property of dynamic epigenes was confirmed, i.e., metastability of some epigenotypes, which was predicted theoretically (Tchuraev, 2006) and found in experiments *in vivo* (Stupak E., Stupak I., 2006). These metastable states realize one of possible mechanisms of primary differentiation of the gene activity patterns that occurs during ontogenesis of multicellular organisms. It is showed on $TCDS^{(-)}$ model, that even systems in which $CDS^{(-)}$ are typical elements, can possess this property.

REFERENCES

Galimzyanov A.V. (2005) Computer modeling the genetic systems controlling ontogenetic processes of eukaryots. *Proc. XVII IMACS World Congress on Sci. Computation, Applied Mathematics and Simulation*, IMACS, Paris, T2-I-47-0310, 1–7.

Hayflick L., Moorhead P.S. (1961) The serial cultivation of human diploid cell strains. *Exp. Cell Res*., **253**, 585–621.

Stupak E.E., Stupak I.V. (2006) Inheritance and state switching of genetic toggle switch in different culture growth phases. *FEMS Microbiology Letters*, **258(**1), 37.

Tchuraev R.N. (1991) A new method for the analysis of the dynamics of the molecular genetic control systems. I. Description of the method of generalized threshold models. *J. Theor. Biol*., **151**, 71–87.

Tchuraev R.N. (2006) General principles of organization and laws of functioning in governing gene networks. In Kolchanov N., Hofestaedt R., Milanesi L. (eds), *Bioinformatics of Genome Regulation and Structure II*. Springer Sci. + Business Media Inc., NY USA, pp. 367–377.

Tchuraev R.N., Stupak I.V., Stupak E.E., Galimzyanov A.V. (2006) A new epigene property: metastable epigenotypes. *Doklady Biol. Sci*., **406**, 97–99.

ASYMMETRIC MODELS OF THE GENE NETWORKS

Golubyatnikov V.P.*[*1], Gaidov Yu.A.[2], Kleshchev A.G.[1]
[1] Institute of Mathematics, SB RAS, Novosibirsk, 630090, Russia; [2] Novosibirsk State University, Novosibirsk, 630090, Russia
[*] Corresponding author: e-mail: glbtn@math.nsc.ru

Key words: dynamical system, mathematical model, negative feedback, closed trajectory, fixed point theorem, Andronov-Hopf bifurcation

SUMMARY

Motivation: Detection of closed trajectories in particular dynamical systems is a hard mathematical problem. We considered it in our previous publications in the case of symmetric dynamical systems as models of the gene networks. It is necessary now to study the general asymmetric gene networks because natural biological processes occur usually in a periodic manner (end-product repression of mRNA synthesis, circadian rhythm, mitotic oscillations, cell cycle engine etc) and do not have any symmetries.

Results: We extend our previous investigations to a much more wide class of the gene network models which are not assumed to be symmetric. We prove existence of periodic trajectories of corresponding dynamical systems. Description of the Andronov-Hopf bifurcation in these models is given as well.

INTRODUCTION

Our main aim is to give a rigorous mathematical explanation of some numerical experiments concerning the limit-cycle oscillations of the gene networks with negative feedbacks described in Likhoshvai *et al*. (2001) and Golubyatnikov *et al.* (2006a, b). In these publications, the right-hand sides of corresponding differential equations were represented by the Hill's functions $F(\bar{x})/G(\bar{x})$ where the variables $\bar{x}$ denote the concentrations of the components, F and G are polynomials with positive coefficients. Most of results obtained there concerned symmetric gene network models. Now, we study more general and more complicated right-hand sides of the equations.

MODEL

We continue our studies (Golubyatnikov *et al*., 2005, 2006a) of special dynamical systems as models of the gene networks regulated by negative feedback control loop on the stages of initiation of the mRNA and/or protein synthesis. Consider 3-dimensional nonlinear dynamical system of a general type:

$$\frac{dx_1}{dt} = f_1(x_3) - x_1;\ \frac{dx_2}{dt} = f_2(x_1) - x_2;\ \frac{dx_3}{dt} = f_3(x_2) - x_3. \qquad (1)$$

Here the functions $f_1(x_3), f_2(x_1), f_3(x_2) : [0,\infty) \to (0,\infty)$ are assumed to be smooth and monotonically decreasing so that $f_i(u) \to 0$, as $u \to \infty$. This system has exactly one stationary point $M_* = (x_{1*}, x_{2*}, x_{3*})$ in the positive octant. Linearization of (1) near its stationary point is described by the matrix

$$A = \begin{pmatrix} -1 & 0 & -p_1 \\ -p_2 & -1 & 0 \\ 0 & -p_3 & -1 \end{pmatrix}, \qquad -p_i = \frac{df_i}{dx_{(i-1)}}(x_{(i-1)*}).$$

We assume here and in the sequel that $x_{(i-1)} = x_3$ for i=1 and denote by P the product $p_1 p_2 p_3$. One of eigenvalues of this matrix A is negative, $\lambda_1 = -1 - \sqrt[3]{P}$ and corresponds to an eigenvector e_1 with positive coordinates. Other its eigenvalues are complex, $2\,\mathrm{Re}\,\lambda_2 = 2\,\mathrm{Re}\,\lambda_3 = \sqrt[3]{P} - 2$, $2\,\mathrm{Im}\,\lambda_{2,3} = \pm\sqrt[3]{P}\sqrt{3}$.

Let $Q = [0, f_1(0)] \times [0, f_2(0)] \times [0, f_3(0)]$. One can verify that this parallelepiped is positively invariant with respect to trajectories of the system (1). Following (Hastings *et al.*, 1977), where quite different types of dynamical systems were studied, consider subdivision $Q = \bigcup S_{ijk}$, i, j, k = 0,1. Here the parallelepipeds S_{ijk} are defined as:

$$S_{000} = \{0 \le x_1 \le x_{1*}; 0 \le x_2 \le x_{2*}; 0 \le x_3 \le x_{3*}\},$$
$$S_{100} = \{x_{1*} \le x_1 \le f_1(0); 0 \le x_2 \le x_{2*}; 0 \le x_3 \le x_{3*}\},$$
$$S_{101} = \{x_{1*} \le x_1 \le f_1(0); 0 \le x_2 \le x_{2*}; x_{3*} \le x_3 \le f_3(0)\} \text{ etc.}$$

Subscript *0* on the *m*-th place means that S_{ijk} borders on the face $x_m = 0$ of Q and subscript 1 on this place means that S_{ijk} borders on the face $x_m = f_{m-1}(0)$ of Q. Note that the vectors $\pm e_1$ show the directions from M_* into the parallelepipeds S_{000} and S_{111}.

RESULTS AND DISCUSSION

1. Denote by F the face $x_2 = x_{2*}$ of the parallelepiped S_{011}. Direct calculations show that trajectories of all points of its face $x_3 = x_{3*}$ enter the parallelepiped S_{010} and that trajectories of all other faces of S_{011} enter S_{011}. Now, consider the parallelepiped S_{010}. As above, for all points of its face $x_1 = x_{1*}$, their trajectories enter the parallelepiped S_{110} and trajectories of the points of all other faces of S_{010} enter S_{010}. After six steps of these shifts along these trajectories $S_{011} \to S_{010} \to S_{110} \to S_{100} \to S_{101} \to S_{001} \to S_{011}$ we see that trajectories of all points of $F \subset S_{001}$ enter S_{011} and trajectories of all points of other faces of S_{001} enter S_{001}, so we get a continuous mapping $\varphi : F \to F$, $F = S_{011} \bigcap S_{001}$.

Let S_6 be the union of all these 6 parallelepipeds above. If P >8, then $\mathrm{Re}\,\lambda_{2,3} > 0$ and the stationary point M_* has a small open neighborhood U such that all points in $U \bigcap S_6$ are moved off M_* during the shifts along the trajectories of (1).

Let $F' = F \setminus (F \bigcap U)$.

This compact contractible set is homeomorphic to disc. Now, φ maps *F'* into *F'*, and the fixed point theorem implies existence of at least one point M_0 in *F'* such that $\varphi(M_0) = M_0$.

Clearly, the trajectory of this point M_0 is a closed cycle and we get

Theorem *If P > 8, then the system (1) has at least one periodic trajectory.*

2. Denote by $W(\bar{x})$ vector field with coordinates $f_1(x_3), f_2(x_1), f_3(x_2)$, so $\frac{d\bar{x}}{dt} = W(\bar{x}) - \bar{x}$ is the vector form of the system (1). Since $div(W(\bar{x}) - \bar{x}) \equiv -3$, the volume of any bounded domain in R^3 decreases during the shifts along the trajectories exponentially: $V(t) = V_0 \cdot e^{-3t}$. Note, that the fixed point theorem and this exponential decreasing do not imply immediately stability or uniqueness of the cycle in our theorem. Nevertheless, numerical experiments with the systems of the type (1) show these stability and uniqueness:

The Fig. 1 demonstrates trajectories convergent to this cycle in the cases of

$f_1(z) = \frac{6}{1+z^5}$; $f_2(x) = \frac{3}{1+x^7}$; $f_3(y) = 7e^{-5y}$, (left);

$f_1(z) = \frac{6}{1+z^5}$; $f_2(x) = \frac{3}{1+x^7}$; $f_3(y) = \frac{7}{1+y^5}$, (right).

The big black points in the centers of the pictures indicate positions of the stationary points of these systems. Similar behavior of trajectories was observed in other numerical experiments with the system (1) and in natural gene networks (Elowitz, Leibner, 2000).

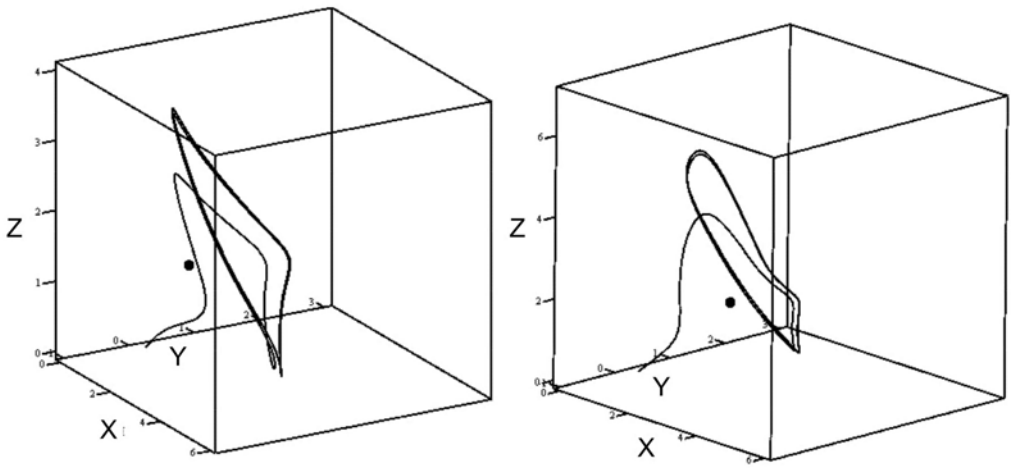


Figure 1. Closed cycles of the system (1).

3. Let α be a real parameter. Consider the dynamical system

$$\frac{dx_1}{dt} = f_1(x_3, \alpha) - x_1;\ \frac{dx_2}{dt} = f_2(x_1, \alpha) - x_2;\ \frac{dx_3}{dt} = f_3(x_2, \alpha) - x_3. \qquad (2)$$

We assume that for all values of α the functions f_i are smooth and monotonically decreasing as in the case of (1). Let $\alpha = \alpha_0$ corresponds to $P = P(\alpha_0) = 8$, this is equivalent to $\mathrm{Re}\lambda_{2,3} = 0$. As in (Golubyatnikov *et al.*, 2006; Volokitin, 2004), the bifurcation theorem implies that if $D_0 \equiv \frac{dP}{d\alpha}(\alpha_0) \neq 0$, and $(\alpha - \alpha_0) \cdot D_0$ is positive and sufficiently small, then the system (2) has a bifurcation cycle in a neighborhood of the point $M_* = M_*(\alpha)$.

4. Analogous results on periodic trajectories and their bifurcations can be obtained for more general type of dynamical systems

$$\frac{dx_1}{dt} = f_1(x_3) - g(x_1);\ \frac{dx_2}{dt} = f_2(x_1) - g(x_2);\ \frac{dx_3}{dt} = f_3(x_2) - g(x_3),$$

where the functions f_i are as above and the function g is monotonically increasing. Similar considerations can be done for other odd-dimensional asymmetric gene network models with the help of sequences φ of the shifts much more complicated than in (Hastings *et al.*, 1977). Biological significance of these mathematical studies and their direct relations to negative feedback oscillations in natural gene networks is discussed in (Elowitz, Leibner, 2000; Likhoshvai *et al.*, 2001).

ACKNOWLEDGEMENTS

The work was supported by leading scientific schools grant No. 8526.2006.1 of President of Russian Federation and by interdisciplinary integration grants NN 24 and 46 of SB RAS. The authors are indebted to E.P. Volokitin for helpful discussions.

REFERENCES

Elowitz M., Leibner S. (2000) A synthetic oscillatory kinetic method for modeling gene network. *Nature*, **403,** 335–339.

Golubyatnikov V.P., Kleshchev A.G., Kleshcheva K.A., Kudryavtzeva A.V. (2006b) Studies of phase portraits of 3-D gene networks models. *Sib. J. of Industrial Mathematics*, **9**(1), 75–84. (In Russ.).

Golubyatnikov V.P., Likhoshvai V.A., Gaidov Yu.A., Kleshchev A.G., Lashina E.A. (2005) Regular and chaotic dynamics in the gene networks modeling. Proc. Intern.l Conf. "Human & Computers-2005", University of Aizu, Japan, 7–12.

Golubyatnikov V.P., Likhoshvai V.A., Volokitin E.P., Gaidov Yu.A., Osipov A.F. (2006a) Periodic Trajectories and Andronov-Hopf Bifurcations in Models of Gene Networks. In Kolchanov N., Hofestaedt R., Milanesi L., (eds), *Bioinformatics of Genome Regulation and Structure II*, Springer Science+Business Media Inc., pp. 405–414.

Hastings S., Tyson J., Webster D. (1977) Existence of periodic solutions for negative feedback cellular control systems. *J. of Diff. Equations*, **25**, 39–64.

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

Volokitin E.P. (2004) On limit cycles in elemental model of hypothetical gene networks. *Sib. J. of Industrial Mathematics*. **7**(3), 57–65. (In Russ.).

GENE NETWORK MODELS WITH DIFFERENT TYPES OF REGULATION

Golubyatnikov V.P.*[*1], Gaidov Yu.A.[2], Kleshchev A.G.[1], Volokitin E.P.[1]
[1] Institute of Mathematics, SB RAS, Novosibirsk, 630090, Russia; [2] Novosibirsk State University, Novosibirsk, 630090, Russia
[*] Corresponding author: e-mail: glbtn@math.nsc.ru

Key words: dynamical system, negative feedback, transcriptional regulation, post-transcriptional regulation, periodic trajectory, fixed point theorem, Andronov-Hopf bifurcation

SUMMARY

Motivation: Determination of periodic trajectories and stationary regimes in gene network models with various types of regulation of their functioning, investigation of their properties are fundamental and challenging problems of bioinformatics.

Results: We prove existence of closed trajectories and describe their bifurcations for a wide class of asymmetric gene network models where the regulation of each of their element is controlled by the post-transcriptional mechanism. These results can be generalized to the case of the gene networks of mixed types, where this regulation is effected either at the stages of initiation of mRNA and proteins synthesis or at the stages of their degradation.

INTRODUCTION

We consider nonlinear dynamical systems as models of gene networks with negative feedbacks which are realized at the stages of degradation of protein synthesis, i.e. are described by the post-transcriptional regulation mechanism. Our previous publications (Golubyatnikov *et al.*, 2005, 2006a) were devoted to the investigations of much more restricted classes of these models, in particular all corresponding dynamical systems were assumed to be symmetric and to have a very special type. Here we continue our studies in a more wide class of gene network models.

IMPLEMENTATIONS AND RESULTS

1. We start with consideration of 3-dimensional nonlinear dynamical system of a general type:

$$\frac{dx_1}{dt} = \alpha_1 - x_1 \cdot g_1(x_3)\ ;\ \frac{dx_2}{dt} = \alpha_2 - x_2 \cdot g_2(x_1)\ ;\ \frac{dx_3}{dt} = \alpha_3 - x_3 \cdot g_3(x_2)\ . \qquad (1)$$

Here $\alpha_j > 0$, $x_j \geq 0$, $j = 1, 2, 3$ and the functions $g_1(x_3)$, $g_2(x_1)$, $g_3(x_2) : [0, \infty) \to [1, \infty)$ are smooth and monotonically increasing so that $g_i(0) = 1$ for all $I = 1, 2, 3$. It follows from the monotone behavior of these functions that the system (1) has exactly one stationary point $M_* = (x_{1*}, x_{2*}, x_{3*})$ in the positive octant. Linearization of this system near its stationary point is described by the matrix with six strictly negative coefficients

$$A=\begin{pmatrix} -s_3 & 0 & -p_1 \\ -p_2 & -s_1 & 0 \\ 0 & -p_3 & -s_2 \end{pmatrix}, \qquad p_i = x_{i*} \cdot \frac{dg_i}{dx_{(i-1)}}(x_{(i-1)*}), \quad s_i = g_i(x_{(i-1)*}).$$

Here and in the sequel, we mean that $x_{i-1} = x_3$ for $I = 1$. Let P be the product $p_1 p_2 p_3$. Since the characteristic polynomial of A has the form $(\lambda + s_1)(\lambda + s_2)(\lambda + s_3) + P = 0$, one of the eigenvalues λ_1 of this matrix A is negative and corresponds to an eigenvector e_1 with positive coordinates. The well-known Routh-Hurwitz criterion gives necessary and sufficient conditions for positivity of the real parts of two other eigenvalues: $\mathrm{Re}\,\lambda_2 = \mathrm{Re}\,\lambda_3 > 0$ if and only if

$$(s_1 + s_2 + s_3)(s_1 s_{2_2} + s_1 s_3 + s_2 s_3) - s_1 s_2 s_3 - P < 0. \qquad (2)$$

Let $Q = [0, \alpha_1] \times [0, \alpha_2] \times [0, \alpha_3]$. This parallelepiped is positively invariant with respect to trajectories of the system (1). As in (Hastings *et al.*, 1977), where quite different types of dynamical systems were studied, consider subdivision $Q = \bigcup S_{ijk}$. Here the parallelepipeds S_{ijk}, *i,j,k = 0* or *1*, are defined as:

$S_{000} = \{0 \le x_1 \le x_{1*};\ 0 \le x_2 \le x_{2*};\ 0 \le x_3 \le x_{3*}\}$,

$S_{101} = \{x_{1*} \le x_1 \le \alpha_1;\ 0 \le x_2 \le x_{2*};\ x_{3*} \le x_3 \le \alpha_3\}$ etc.

The subscript *0* on the *m*-th place means that S_{ijk} borders on the face $x_m = 0$ of Q, subscript 1 on this place means that S_{ijk} borders on the face $x_m = \alpha_m$ of Q and the vectors $\pm e_1$ show the directions from the point M_* into the parallelepipeds S_{000} and S_{111}. Now, our arguments in this paragraph will follow almost literally (Golubyatnikov *et al.*, 2006a). The shifts $S_{011} \to S_{010} \to S_{110} \to S_{100} \to S_{101} \to S_{001} \to S_{011}$ along trajectories of the system (1) determine a continuous mapping $\varphi : F \to F$ of the rectangle $F = S_{011} \bigcap S_{001}$. If the inequality (2) is satisfied, then the fixed point theorem implies existence of at least one point M_0 in $F' = F \setminus (F \bigcap U)$ such that $\varphi(M_0) = M_0$. Here U is some small open neighborhood of the point M_0. So, the trajectory of this point M_0 is a closed cycle and we obtain

Theorem. *If the condition (2) is satisfied, then the system (1) has at least one periodic trajectory.*

2. Let $(\alpha_1 - x_1 \cdot g_1(x_3); \alpha_2 - x_2 \cdot g_2(x_1);\ \alpha_3 - x_3 \cdot g_3(x_2))$ be the coordinates of vector field $G(\bar{x})$. Hence, $\frac{d\bar{x}}{dt} = G(\bar{x})$ is the vector form of the system (1). We see that $div(G(\bar{x})) \le -3$, so, the volume of any bounded domain in R^3 decreases during the shifts along the trajectories exponentially: $V(t) < V_0 \cdot e^{-3t}$. Again, the fixed point theorem and this decreasing do not imply immediately stability or uniqueness of the cycle in our

theorem, though numerical experiments with this system and with more complicated systems show that this cycle is stable and unique, as in (Golubyatnikov *et al.*, 2006a, b) and below.

3. Let μ be a real parameter. Consider the dynamical system

$$\frac{dx_1}{dt} = \alpha_1(\mu) - x_1 g_1(x_3, \mu);\ \frac{dx_2}{dt} = \alpha_2(\mu) - x_2 g_2(x_1, \mu);\ \frac{dx_3}{dt} = \alpha_3(\mu) - x_3 g_3(x_2, \mu)\ . \qquad (3)$$

We assume that for all values of μ the functions g_i are smooth and monotonically increasing as in the case of (1). Let $\mu = \mu_0$ corresponds to $\mathrm{Re}\lambda_{2,3} = 0$. As in our previous considerations (Volokitin, Treskov, 2005; Golubyatnikov *et al.*, 2006a), the Hopf bifurcation theorem, see for example (Kuznetzov, 1995), implies that if $(\mu - \mu_0) \cdot \frac{dP}{d\mu}(\mu_0)$ is positive and sufficiently small, then the system (3) has a bifurcation cycle in a neighborhood of the point $M_* = M_*(\mu)$. In the case of symmetric gene networks of special type, we have obtained the stability conditions of these bifurcation cycles based on the formula for Lyapunov parameter ν_1. Similar explicit formulae can be derived for the asymmetric system (3) as well, but in this case the analytical expression of ν_1 is too cumbersome and can be used only in the numerical experiments.

4. Using exactly the same arguments, analogous results on periodic trajectories and their bifurcations can be obtained for the gene networks of mixed types, where some of the equations have the form (1) and other equations describe regulation on the stages of the genes expression activation, i.e., have the form $\frac{dx_i}{dt} = f_i(x_{i-1}) - x_i$. Here, the functions f_i are monotonically decreasing as in (Golubyatnikov *et al.*, 2006b).

The left and the right parts of the Fig. 1 show (respectively) trajectories convergent to the bifurcation cycles in dynamical systems

$$\frac{dx}{dt} = \frac{\alpha_1}{1+z^3} - x;\ \frac{dy}{dt} = \frac{9}{1+x^3} - y;\ \frac{dz}{dt} = \alpha_3 - z(1+y^3);\ \alpha_1 = 8.02,\ \alpha_3 = 2.88,$$

$$\frac{dx}{dt} = \alpha_1 - x(1+z^3);\ \frac{dy}{dt} = \frac{9}{1+x^3} - y;\ \frac{dz}{dt} = \alpha_3 - z(1+y^3);\ \alpha_1 = 6.15,\ \alpha_3 = 2.4.$$

The black "eyes" on the pictures indicate positions of the stationary points of these systems.

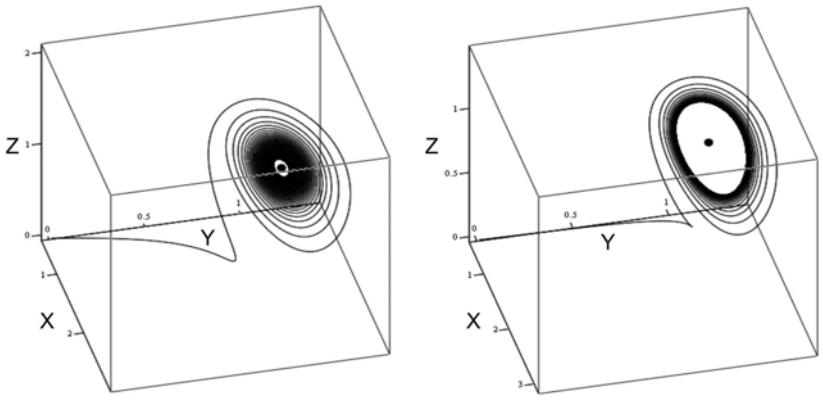


Figure 1. Andronov-Hopf bifurcations in gene networks models of mixed types.

DISCUSSION

Separation of regular and chaotic domains in the spaces of parameters is an actual problem of the gene network modeling. In some higher-dimensional symmetric gene network models, we have already discovered appearance of chaotic behavior of trajectories (Golubyatnikov *et al*., 2005). Similar behavior of trajectories was observed in natural gene networks functioning (Gardner *et al.*, 2000), and in other numerical experiments with the systems considered above and in (Golubyatnikov *et al.*, 2006b). So, one of our current tasks is to determine conditions of regular behavior of these trajectories and to study questions of stability and uniqueness of the cycles discovered in the dynamical systems considered above, bifurcations of these cycles and their higher-dimensional analogues.

ACKNOWLEDGEMENTS

The work was supported by leading scientific schools grant No. 8526.2006.1 of President of Russian Federation and by interdisciplinary integration grants 24 and 46 of SB RAS. The authors are indebted to V.A. Churkin for discussions.

REFERENCES

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

Golubyatnikov V.P., Gaidov Yu.A., Kleshchev A.G. (2006a) Asymmetric gene networks. *This issue*.

Golubyatnikov V.P., Likhoshvai V.A., Gaidov Yu.A., Kleshchev A.G., Lashina E.A. (2005) Regular and chaotic dynamics in the gene networks modeling. Proc. Intern. Conf. "Human & Computers-2005", University of Aizu, Japan, 7–12.

Golubyatnikov V.P., Likhoshvai V.A., Volokitin E.P., Gaidov Yu.A., Osipov A.F. (2006a) Periodic Trajectories and Andronov-Hopf Bifurcations in Models of Gene Networks. In Kolchanov N., Hofestaedt R., Milanesi L. (eds), *Bioinformatics of Genome Regulation and Structure II*, Springer Science+Business Media Inc., pp. 405–414.

Hastings S., Tyson J., Webster D. (1977) Existence of periodic solutions for negative feedback cellular control systems. *J. of Diff. Equations*, **25**, 39–64.

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

Volokitin E.P., Treskov S.A. (2005) Andronov-Hopf bifurcation in model of hypothetical gene networks. *Sib. J. of Industrial Mathematics*. **8(**1), 30–40. (In Russ.).

RESEARCH OF CYCLIC GENE NETWORK CIRCUITS WITH NEGATIVE TYPE OF REGULATION

***Klishevich M.A.*[*1], *Kogai V.V.*[1], *Fadeev S.I.*[1]**
[1] Sobolev Institute of Mathematics, SB RAS, Novosibirsk, 630090, Russia
[*] Corresponding author: e-mail: mans1@ngs.ru

Key words: gene network, periodic solutions, parameter continuation, delay argument equation, boundary value problem

SUMMARY

Motivation: The core of gene networks (GN) is regulatory circuits – genes and proteins with mutually regulated expression. Thus, detection of possible operation modes of regulatory circuits is an important problem of systems biology.

Results: The work presents an efficient method for computing symmetric periodic solutions for cyclic regulatory circuits for model M(*n*,2). This method includes resolving boundary value problem for equation with delay argument (2) and receiving functional relation T/τ by means of parameter continuation.

INTRODUCTION

Gene networks (GN) are structurally complex spatial objects composed of hundreds of elements of various natures and complexities, namely, genes and their regulatory regions, RNAs and proteins encoded by these genes, low-molecular-weight compounds, various complexes between enzymes and their targets, etc. The core of GN is regulatory circuits—genes and proteins with mutually regulated expression. Their presence confers on GN a unique ability to respond adequately to changes in external conditions. Thus, detection of possible operation modes of regulatory circuits is important problem of systematic biology. A constructive step in this direction is separation of a finite set of standard elements from natural GN, formalization of the rules for assembling theoretical objects (mathematical models) describing regulatory circuits from these elements, and a systematic analysis of their properties for revealing general biologically significant regularities (Likhoshvai *et al*., 2003).

In this work we study a model describing the most simple cyclic circuits of molecular-genetic systems.

Mathematical model is presented by autonomous system of n differential equations:

$$\frac{dx_i}{dt} = \frac{\alpha}{1 + x_{i-1}^{\gamma}} - x_i, i = \overline{1,n}, \text{ let } i-1 = 0, \text{ let } i-1 = n. \qquad (1)$$

Here, *n* is number of gene elements in our system. Where *i*-th gene element, *I* = 1, …, n, encodes its protein-regulator. The value of *i*-th component means a concentration of protein has been synthesized as a result of *i*-th gene element expression. The positive term in the right part of *i*-th equation describes a negative type regulatory effect of previous gene element to the efficacy of expression of current element. α, γ are positive parameters, which set level of basal activity of expression and nonlinearity of regulator

effect. All negative type regulation mechanisms consider being identical in the model. Negative term describes degradation process. System is regarded to be dimensionless, that is why time is relative. Those facts were shown before: (i) phase trajectories of system (1) coming from the hypercube Ξ with the edge α remain in Ξ, (ii) system (1) always has one stationary symmetric point, which is stable and single on the assumption of α, γ – are sufficiently small, (iii) subject to α, γ are sufficiently great the symmetric point lose its stability, (iv) if n is even two partially symmetric stable stationary points arise in the system, (v) if α, γ are sufficiently great oscillation trajectories which are symmetric appear to exist, (iv) in order to answer the question about their stability we use $(n,k)-$ criterion. The question about quantity of symmetric cycles of (1) is to be examined. Given work presents an efficient method for searching symmetric cycles for system (1), so-called Model M(n,2). The computation is based on the analyses of equation (2), which by itself is a model of auto-repressilator, single gene element able to repress its own expression (Likhoshvai *et al.*, 2005).

METHODS AND ALGORITHMS

Consider the boundary value problem for delay equation that naturally appears from the problem of finding symmetric periodic solutions of system (1). Such solution consists of n identical periodic trajectories, which are shifted for value τ from each other.

$$\frac{du(t)}{dt} = T\left(\frac{\alpha}{1+u^{\gamma}(t-\tau)} - u(t)\right),\ u(0)-u(1)=0,\ \ u(0)-\frac{\alpha}{1+u^{\gamma}(-\tau)}=0. \qquad (2)$$

Investigation of problem (2) is connected with approximate discrete model representation, that appears to be the system of n nonlinear equations for the mesh values of function $u(t)$ (Fadeev, 1990).

For this purpose we integrate both parts of differential equation (2) on interval $[t_i, t_{i+1}]$:

$$u_{i+1}-u_i = \alpha T\int_{t_i}^{t_{i+1}} f(u(t-\tau))dt - T\int_{t_i}^{t_{i+1}} u(t)dt, i=1,...,N-1,$$

Lets denote:

$$F_i \equiv u_{i+1}-u_i + T\int_{t_i}^{t_{i+1}} u(t)dt - \alpha T\int_{t_i}^{t_{i+1}} f(u(t-\tau))dt = 0, i=1,...,N-1, \qquad (3)$$

where T – period, τ – delay, $f(x) = 1\Big/\left(1+x^{\gamma}\right)$.

Integrals in (3) are approximately calculated using Hermite parabola

$$\int_{t_i}^{t_{i+1}} u(t)dt = \frac{h_i}{2}(u_i+u_{i+1}) + \frac{h_i^2}{12}(u_i^{'}-u_{i+1}^{'}),$$

and Simpson quadrature formula

$$\int_{t_i}^{t_{i+1}} f\left(u(t-\tau)\right)dt = \frac{h_i}{6}\left(f\left(u(t_i-\tau)\right) + 4f\left(u\left(\frac{t_i+t_{i+1}}{2}-\tau\right)\right) + f\left(u(t_{i+1}-\tau)\right)\right).$$

For system (3) we add periodicity condition – second equation (2), and transversality condition – third equation (2). Finally we obtain the system of non-linear rational equations which we solve by Newton method.

We supply our numerical method with mesh adaptation, it holds up computational accuracy for great gradient domains. The adaptation is based on the analytical boundary value problem (1) solution for $\gamma = \infty$. The limiting solution, that is

$$u(t) = \begin{cases} (1-\alpha)\exp(t_1 - t) + \alpha, t \in [0, t_2], \\ u_2 \exp(t_2 - t), t \in [t_2, T], \end{cases}$$

almost coincide with the solution for (2) for γ sufficiently great.

IMPLEMENTATION AND RESULTS

Using parameter T continuation for boundary value problem (2) we may calculate functional dependence T/τ from τ (Fig. 1) and thus to obtain plurality of solutions for model M(n,2) in the domain of parameter modification (Fadeev *et al.*, 1998).

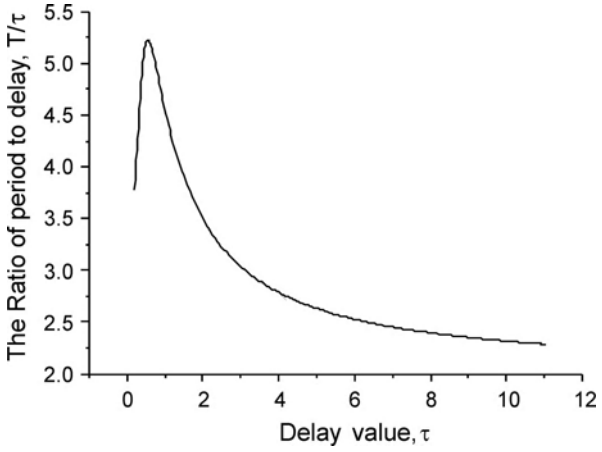


Figure 1. Dependence of the ratio $\frac{T}{\tau}$ from τ, $\alpha = 25$, $\gamma = 10$.

Fig. 1 shows dependence T/τ from τ, according to it, we may conclude that, for example, model M(5,2) has 3 solutions. Components of symmetric periodic solution for model M(n,2) may differ one from the other for the constant value τ, where by virtue of symmetry τ must be multiple $1/T$ and satisfy inequality $0 < \tau < 1$. That is why the only values T/τ receive are 5 and 5/2. For this magnitudes correspond τ values 6.3313, 0.7503 and 0.4181.

For those τ we compute the solution of problem (2).

Fig. 2 demonstrates periodic solution for $\tau = 6.3313$. Chart presents dependence $u(t)$, where time is T-normalized.

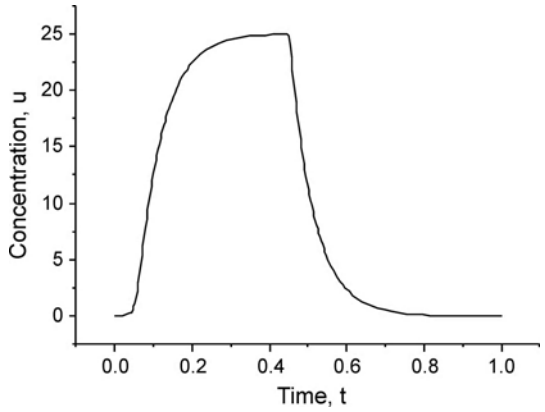


Figure 2. Solution $u(t)$ for (2), $\alpha = 25$, $\gamma = 10$, $\tau = 6.3313$, $T = 15.7331$. Time is relative to period T.

DISCUSSION

Studying properties of model M(n,2) appears to be an important problem for the theory of gene network. Particularly, this models describe for $n = 2$ molecular trigger and for $n = 3$ repressilator, such constructions were realized by gene-engineering methods (Gardner *et al.*, 2000; Elowitz, Leibler, 2000; Tchuraev *et al.*, 2000). The model have been studied is interesting as well for more complicated molecular-genetic systems construction (Sprinzak, Elowitz, 2005). That is why it is important to reveal in simple regulatory circuits all types of dynamic behavior.

Method have been suggested allows to find efficiently symmetric periodic trajectories in gene network mathematical models class M(n,2). Let us note, that all symmetric model M(n,2) cycles for arbitrary n are cycles of auto-repressilator model. However, question about complicated types of periodical trajectories is not already examined. Computational investigation shows, that for even n partially-symmetric periodical trajectories may be found for model M(n,2). All components of those solutions are divided into two groups; in each group trajectories are identical and shifted from each other for some phase. All such periodic solutions satisfy the system of two equations with delay:

$$\frac{dy_1}{dt} = \frac{\alpha}{1 + y_2^{\gamma}(t-\tau)} - y_1, \quad \frac{dy_2}{dt} = \frac{\alpha}{1 + y_1^{\gamma}(t-\tau)} - y_2,$$

which present a molecular trigger model, that keeps in mind times of protein-regulator synthesis. Consequently, we could expect molecular trigger to have nonsymmetrical cycle. That is why transient regime from one stationary point to another (switching of molecular trigger) could go through oscillation trajectory, twisting on that cycle arbitrary long. This trajectory is periodic-like. Since the cycle is not stable, fluctuations sooner or later will transfer it to stable regime. In future we are to examine oscillation parameters for more complicated models of molecular trigger in order to appreciate possibility of its existence naturally.

ACKNOWLEDGEMENTS

Work was supported in part by the Russian Foundation for Basic Research (No. 06-04-49556). The authors are grateful to Vitaly Likhoshvai for valuable discussions.

REFERENCES

Elowitz M.B., Leibler S. (2000) A synthetic oscillatory network of transcriptional regulators. *Nature*, **403**(6767), 335–338.

Fadeev S.I. (1990) Program for numerical non-linear boundary value problems resolving for ordinary differential equation systems with parameter. In Godunov S.K., (ed.), *The Computational Methods of Linear Algebra. Novosibirsk: Science, SB: The Proceedings of Sobolev Institute of Mathematics SB RAS*. **17**, pp. 104–198. (In Russ.).

Fadeev S.I., Pokrovskaya S.A., Berezin A.Yu., Gainova I.A. (1998) Software Package STEP for Numerical Study of General Systems of Nonlinear Equations and Autonomous Systems. *Novosibirsk, Novosibirsk State University Press*. (In Russ.).

Gardner T.S., Cantor C.R., Collins J.J. (2000) Construction of a genetic toggle switch in *Escherichia coli*. *Nature*, **403**(6767), 339–342.

Likhoshvai V.A., Kogai V.V., Fadeev S.I. (2005) Self-Oscillations in Hypothetical Gene Networks. In Kolchanov N., Hofestaedt R., (eds), *Bioinformatics of Genome Regulation and Structure II*. Springer Science + Business Media, Inc. 391–404.

Likhoshvai V.A., Matushkin Yu.G., Fadeev S.I. (2003) Problems of the theory of gene network operation. *Sib. Zh. Industr. Matem*., **4,** 64–80 (In Russ.).

Sprinzak D., Elowitz M.B. (2005) Reconstruction of genetic circuits. *Nature*, **438**, 443–448.

Tchuraev R.N., Stupak I.V., Tropynina T.S., Stupak E.E. (2000) Epigenes: design and construction of new hereditary units. *FEBS Lett*., **486**(3), 200–202.

ON THE RECONSTRUCTION OF A GENETIC AUTOMATON ON THE BASIS OF BOOLEAN DYNAMIC DATA

***Komarov A.V.*[*1], *Akberdin I.R.*[2], *Ozonov E.A.*[1], *Evdokimov A.A.*[3]**
[1] Novosibirsk State University, Novosibirsk, 630090, Russia; [2] Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia; [3] Sobolev Institutes of Mathematics, SB RAS, Novosibirsk, 630090, Russia
* Corresponding author: e-mail: medium@gorodok.net

Key words: genetic automaton, regulatory pattern of gene network, mechanisms of expression gene regulations, Boolean function, reconstruction and implicants

SUMMARY

Motivation: The reconstruction of connections and mechanisms of interactions between genes in a regulatory profile of gene networks on basis of different types of experimental data is the most important problem in system biology and bioinformatics. In fact, this encompasses the development of mathematical models of interactions between genes and reconstruction algorithms for these mechanisms.

Results: Reconstruction algorithm of Boolean functions for subclass models of genetic automata is proposed.

INTRODUCTION

Cell phenotypes are determined by the concerted activity of thousands of genes and their products. This activity is coordinated by a complex network, which regulates the expression of genes controlling common functions, such as the formation of a transcriptional complex or the availability of a signaling pathway. Understanding this organization is crucial to elucidate normal cell physiology as well as to dissect complex pathologic phenotypes (Basso *et al.*, 2005).

The problem of the reconstruction of gene networks by computational methods on the basis of different types of dynamic experimental data has increasingly gaining importance with the advent and progressive development of technologies, which allowed measuring the expression of a number of genes together. At present, this problem can be conditionally divided into three sub problems: 1) network topology (reconstruction of undirected connections); 2) reconstruction of connections between genes and their types (activation, inhibition); 3) creation of mathematical models that describe the dynamics of the gene network.

Structure of gene networks and adequate mathematical models of such networks are very useful for prediction of effects of drugs and mutations on the organism.

In this work, we are developing an approach for reconstruction of gene regulation mechanisms given that the oriented arcs in gene network graphs are known. A genetic automaton with zero thresholds is used as the model of gene expression regulation. The problem is to reconstruct a partially defined Boolean function in a special form.

ALGORITHMS AND METHODS

A genetic automaton with zero thresholds is used as the model of the regulatory profile of a gene network (Likhoshvai, 2006).

A partial problem being addressed is one of reconstructing the unknown regulatory mechanisms of the activity of an individual element of the genetic automaton with zero thresholds. It is assumed that the number of input signals is known and so is the response of the mechanism to a certain spectrum of input signals. It is assumed that the thresholds of the logical functions are equal to zero. These functions describe the regulatory mechanisms of the activities of elements of the genetic automaton. Given that, the input variables of the logical function data unambiguously become Boolean (0 for zero concentrations, 1 for nonzero concentrations). The regulatory mechanism is a Boolean function in a special form as follows:

$$B = \bigvee_{i:\delta_i=0} \left(\bigwedge_{k=1}^{l(i)} x_{i_k}\right) \vee \left(\bigwedge_{i:\delta_i=1} \overline{\left(\bigwedge_{k=1}^{l(i)} x_{i_k}\right)}\right). \quad (1)$$

Where index δ_i denotes i-th mechanism of threshold inhibition if $\delta_i = 0$ or i-th mechanism threshold activation otherwise. All inputs of each mechanism are ordered and has internal index from 1 to l(i). So index i_k means number of input in this internal order of i-th mechanism. Boolean formula (1) gives the collective regulation of target element by inputs. For more details, please refer (Likhoshvai, 2006).

After using rules of de Morgan and performing algebraic transformations, we can obtain the Boolean function in disjoint normal forms (DNF)

$$B = \bigvee_{i:\delta_i=0} \left(\bigwedge_{k=1}^{l(i)} x_{i_k}\right) \vee \left(\bigvee_{\substack{j_1\in(1,..,l(1)) \\ \cdots \\ j_r\in(1,..,l(r))}} \left(\bigwedge_{\substack{i=1 \\ i:\delta_i=1}}^{r} \overline{x_{i,j_i}}\right)\right). \quad (2)$$

Assume without loss of generality that index i varies from 1 to r for activation mechanisms then DNF of formula (1) looks like (2).

Suppose the experiment yielded a number of variables (x_1, …, x_n) of the Boolean function B such that it takes on 0 or 1, respectively (Table 1). It should be noticed that one might not have obtained values of Boolean function B on all possible binary vectors. So often we have partially defined Boolean function because we don't know function value on all their inputs, i.e. their true table is incomplete. Obviously, function B is true either on the vector (1,1, …, 1) or on the (0,0, …, 0) in accordance with formula (2). It is necessary to find the form of partially defined Boolean functions as the one in Eq. (2). For the description of the algorithms, we will stick to definitions given by (Sholomov, 1980): The implicant of the Boolean function f (x_1, …, x_n) is such a Boolean function g(x_1, …, x_n) that $g = g \wedge f$. Say that conjunction K covers some binary vector as (x_1, …, x_n), if and only if K is true on it. For example, K = $x_2 \cdot \overline{x_3}$ covers the vector (1,1,0) and does not the vector (1,0,0). Call conjunction K the simple implicant for a partially defined Boolean function is this conjunction:1) covers at least one vector (x_1, …, x_n) in T_1; 2) does not cover any vector (x_1, …, x_n) in T_0; 3) any conjunction derived from it by elimination of variables covers a vector (x_1, …, x_n) in T_0. Obviously, the partially defined Boolean function *f* can be presented as the disjunction of its simple implicants, which in the general case can be done in more than one way. Call the disjunction of simple impicants that is true on the T_1 set and false on the T_0 set the coverage of the function *f*. The main idea of the proposed method is a search for such special-form implicants of a partially

defined Boolean function that their disjunction allows the function to assume the form as in Eq. (2).

The algorithm consists of the following steps.

Step 1. Form a table of the experimental vectors for the values of input parameters with which the function B takes value 0 and 1 (in the table: columns T_0 and T_1). Set $T_3 = \varnothing$.

Step 2. Form the vectors set T_2 from the vector set T_1. The set T_2 includes all possible vectors in the set T_1 that formed by elimination of either 0 or 1 and do not match any vector in T_0 at not "strikeout" positions. If T_2 contains no vector, assume $T_1 = T_2$ and go to Step 5. Vectors in subset T_2 are represented in the {0,1,*} alphabet where symbol * corresponds to "strikeout" position. The words of length *n* in this code correspond to the conjunctions which include the i-th variable with negation if there is 0 at position i, include the i-th variable without negation if there is 1 at position i, and don't include the i-th variable if there is * at position i. These conjunctions are true on some of the Boolean vectors from T_1 and false on any vector from T_0.

Step 3. Label all the vectors in T_1 that are covered by conjunctions corresponding to the vectors in T_2. Vectors that remain unlabelled are to be included in T_3.

Step 4. Assume $T_1 = T_2$, $T_2 = \varnothing$. Go to Step 2.

Step 5. Form the set $T = T_2 \cup T_3$.

Step 6. Eliminate vectors containing both 0 and 1 from T. If T is empty, the experimental data do not permit the Boolean function to assume the form as in Eq. (2).

Step 7. Compose DNF for each subset T that covers the function *f*. If none does so, the data not permit the Boolean function to assume the form as in Eq. (2).

Step 8. Decompose the disjunctions which include input variables with negation into the following conjunction: $\bigwedge\limits_{i:\delta_i=1} \overline{(\bigwedge\limits_{k=1}^{l(i)} x_{i_k})}$. Obtain the Boolean function in the form as in Eq. (1), which is the desired mechanism.

RESULTS

Consider how the algorithm works with the following example: Let the regulatory mechanism of expression of the genetic element *y* (Fig. 1) is represented as the following Boolean formula:

$$y = B(x_1, x_2, x_3, x_4) = x_1 \cdot x_2 \vee \overline{x_2} \cdot \overline{x_3 \cdot x_4}. \qquad (3)$$

Experimental observations of $B(x_1, x_2, x_3, x_4)$ behaviors presented in Table 1 are simulated by the mechanism that presented in Fig. 1. To demonstrate the results of the algorithm, the values of the vectors in T_2 and T at different steps are presented in Table 2. At step 5, the set T is formed that corresponds to the conjunctions $\{\overline{x_2} \cdot \overline{x_3},\ \overline{x_1} \cdot \overline{x_2},\ x_3 \cdot \overline{x_4},\ \overline{x_2} \cdot \overline{x_4},\ \overline{x_1} \cdot x_3,\ x_1 \cdot \overline{x_4},\ x_1 \cdot \overline{x_3},\ x_1 \cdot x_2\}$, of which only $P = \{x_1 \cdot x_2, \overline{x_2} \cdot \overline{x_3},\ \overline{x_1} \cdot \overline{x_2},\ \overline{x_2} \cdot \overline{x_4}\}$ conform to the form as in Eq. (2). The coverage of the table-defined function by implicants in P is possible in two ways: $\{x_1 \cdot x_2 \vee \overline{x_2} \cdot \overline{x_4} \vee \overline{x_2} \cdot \overline{x_3},\ x_1 \cdot x_2 \vee \overline{x_2} \cdot \overline{x_4} \vee \overline{x_1} \cdot \overline{x_2}\}$. These two functions correspond to two different mechanisms of regulation ($x_1 \cdot x_2 \vee \overline{x_2} \cdot \overline{x_3 \cdot x_4}$, $x_1 \cdot x_2 \vee \overline{x_2} \cdot \overline{x_1 \cdot x_4}$): one, which is required, and another, which is presented in Fig. 2.

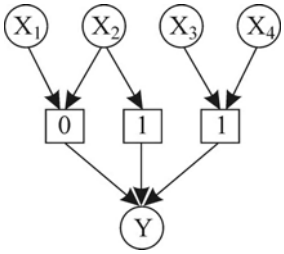


Figure 1. The regulation mechanism used for simulating experimental data.

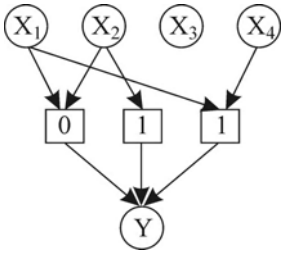


Figure 2. One of the mechanisms proposed by the algorithm.

Table 1. The values of the parameters for which the response from the function B is known

T_0: B=0	T_1: B=1
0100, 0101, 1011	0001, 0010, 1000, 1010, 1100

Table 2. The values of the parameters in sets T2 and T at different steps of the algorithm

T2 after the first instance of step 2	T = T2 after the second instance of step 2	Set T after step 6
*001, 00*1, 000*, 0*10, 00*0, 001*, *000, 10*0, 100*, *010, 1*10, 10*0, 1*00, 11*0, 110*	*00*, 00**, **10, *0*0, 0*1*, 1**0, 1*0*, 11**	*00*, 00**, *0*0, 11**

DISCUSSION

The creation of the gene automaton, which reconstructs the mechanisms of gene regulation, given that the directed arcs in the gene network graph are already known, is the next important step in studying regulatory profiles of gene networks. It allows not only significant interactions between genes to be reconstructed but also the regulatory mechanisms of their interactions to be determined, which provides an important insight into post genomic studies. However, the usability of the algorithm is limited by the creation of effective reliable and universally recognized methods for reconstruction of the regulatory profile of the gene network. In this work, a Boolean function-based algorithm for the reconstruction of regulatory mechanisms in gene networks has been proposed and implemented. Although gene automata with Boolean functions as regulatory mechanism look very similar to the Boolean networks actively studied in elsewhere (Thomas *et al.*, 1995; Akutsu *et al.*, 1998–2000; Liang *et al.*, 1998), their one clear-cut difference. Gene automaton dynamics is described by piecewise linear functions, which have a potential to describe experimental data more accurately than Boolean networks. In the case of non-

zero thresholds regulatory mechanisms have more complex structure then Boolean function and well tractable biologically (Likhoshvai, 2006). The algorithm allows the reconstruction of mechanisms with quite a large number of regulators to be performed in the settings of insufficient experimental data. The algorithm reconstructs several variants of regulatory profiles of the gene network being studied, which does not contradict to a possible availability of several regulatory variants that yield identical responses. However, since the proposed algorithm reconstructs several variants of regulatory mechanisms, the next step in gene network reconstruction is an expert's choice of a biologically more feasible selection from among all possible mechanisms. We are planning to extend the algorithm to gene automata with non-zero thresholds and to integrate the algorithm with methods for the reconstruction of structural connections and with algorithms for the solution of the inverse parametric problem with a view to the reconstruction of regulatory mechanisms using Hill's generalized functions. We are planning to develop methods for setting up new experiments to deliver the best performance of the reconstruction of regulatory mechanisms.

ACKNOWLEDGEMENTS

Work was supported in part by the Russian Foundation for Basic Research (No. 06-04-49556). The authors are grateful to Vitaly Likhoshvai for valuable discussions.

REFERENCES

Akutsu T., Kuhara S., Maruyama O., Miyano S. (1998) Identification of gene regulatory networks by strategic gene disruptions and gene overexpressions. *Proceedings of 9th ACM-SIAM Symposium Discrete Algorithms*, 695.

Akutsu T., Miyano S., Kuhara S. (1999) Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. *Proceedings of Pacific Symposium on Biocomputing'99 (PSB'99)*, 17–28.

Akutsu T., Miyano S., Kuhara S. (2000) Inferring qualitative relations in genetic networks and metabolic pathways. *Bioinformatics*, **16**, 727–734.

Basso K., Margolin A., Stolovitzky G., Klein U., Dalla-Favera R., Califano A. (2005) Reverse engineering of regulatory networks in human B cells. *Nature Genetics*, **37**, 382–390.

Liang S., Fuhrman S., Somogyi R. (1998) Reveal, a general reverse engineering algorithm for inference of genetic network architectures. *Proceedings of Pacific Symposium on Biocomputing (PSB'98)*, 18–29.

Likhoshvai V.A. (2006) On the problem of search for stationary points in regulatory circuits of gene networks. In Kolchanov N., Hofestaedt R., Milanesi L. (eds), *Bioinformatics of Genome Regulation and Structure II.* Springer Science+business Media, pp. 415–420.

Sholomov L.A. (1980) *Bases of theory discrete logical and calculating devices.* Moscow, Science, pp. 65–85.

Thomas R., Thieffry D., Kaufman M. (1995) Dynamical behavior of biological regulatory networks. Biological role of feedback loops and practical use of the concept of the loop-characteristic state. *Bull. of Mathematical Biol.*, **57**, 247.

OSCILLATIONS OF CHAOTIC TYPE IN SYMMETRIC GENE NETWORKS OF SMALL DIMENSION

***Likhoshvai V.A.*[*1], *Rudneva D.S.*[2], *Fadeev S.I.*[3]**
[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia; [2] Novosibirsk State University, Novosibirsk, 630090, Russia; [3] Sobolev Institute of Mathematics, SB RAS, Novosibirsk, 630090, Russia
[*] Corresponding author: e-mail: likho@bionet.nsc.ru

Key words: gene networks, autonomous systems, symmetry, deterministic chaos

SUMMARY

Motivation: Study of functioning laws of regulatory circuits of gene networks is one of the actual problems of mathematical biology.

Results: A mathematical model of a symmetric genetic construction consisting of three gene elements is considered. We show that the model can have aperiodic functioning regimes of chaotic type in spite of compete symmetry of the system with respect to permutation of variables.

Availability: Available on request.

INTRODUCTION

One of the actual problems of modern systemic biology is study of functioning laws of regulatory circuits of gene networks. In particular, we have the significant problem of research of relationships between structures of regulatory connections and kinds of mechanisms and dynamic characteristics of behavior of gene networks. In the framework of the problem it is important to study dynamic behavior of gene networks consisting of a small number (at most three) of identical genetic elements. Such gene networks can be simulated by systems of the form

$$\frac{dx_i}{dt} \quad f(z_i(t \quad)) \quad x_i, \qquad i \quad \overline{1,n.} \tag{1}$$

Here $z_i \quad {}^{n}_{j\ 1} {}_{ij}x_j, \delta_*$ are logical parameters equaled to 0,1, f is a function describing the mechanism of expression regulation of the genetic element, τ is a delay characterizing matrix processes in biological systems. Presence of matrix processes is one of the important factors of gene networks because, together with the mechanism of genetic regulation of gene expression, it can be the cause of cyclic and more complicated (chaotic) types of dynamic behavior of genetic systems consisting only of one genetic element (Mackey, Glass, 1977). Therefore, in the general case, elimination of the delay from the system (1) bring into simplification of dynamics of their behavior. However, if n is odd and greater that 1 a complicated behavior can be obtained in the absence of the delay.

In this paper we propose a mathematical model of a symmetric molecular genetics system consisting of three identical genetic elements (n = 3). The model can be represented by the autonomous system of equations

$$\frac{dx_1}{dt} = P^{(m)}(u_1) - x_1 \ , \ \frac{dx_2}{dt} = P^{(m)}(u_2) - x_2 \ , \ \frac{dx_3}{dt} = P^{(m)}(u_3) - x_3 \ , \ m = 1, 2, \ldots . , \quad (2)$$

where

$$u_1 = x_2 + x_3, \quad u_2 = x_3 + x_1, \quad u_3 = x_1 + x_2,$$

$$P^{(m)}(z) = \alpha \frac{P^{(m-1)}(z)}{1 + [P^{(m-1)}(z)]^{\gamma}}, \ m = 2, 3, \ldots . , \quad P^{(1)}(z) = \frac{\alpha z}{1 + z^{\gamma}} .$$

We study behavior of limit solutions of the system with respect to the parameter α, meaning it limiting decisions the difference scheme used at integration of system (2). We propose a method to indicate domains of α where the system has complicated aperiodic behavior.

METHODS

We study behavior of the model (2) with respect to the parameter α by the method of evolutionary diagrams (Akhromeeva et al., 1988). In the present paper two types of diagrams are proposed. Diagram of the first type (D1) displays properties of solutions of the autonomous system (2) with respect to discrete slowly varying values of the parameter α. For each α graphs of the components $x_1(t), x_2(t), x_3(t)$ are projected on the straight line $\alpha = const$ of the (α, y)-plane from values of t under which the graphs give limit solutions. Diagram of the second type (D2) plays the role of a filter that "filters out" limits solutions such as stable stationary solutions and limit cycles and keeps only oscillations of complicated structures. To achieve this purpose, on the straight line $\alpha = const$ we project mappings of solutions onto the Poincare plane. Another variant of the filter (D3) uses the distance between solutions of (2) with undisturbed and slightly disturbed initial data.

RESULTS

Using the method of constructing D3, we define all partially symmetric and nonsymmetric stationary solutions and analogous limit cycles for the system (2). Then, using the package STEP (Fadeev *et al*., 1998), the obtained stationary solutions are taken as initial solutions for the parameter continuation method in order to study the dependence on the parameter α and the stability. Using the methods of constructing D2 and D3, for m = 3, m = 4 and different values of γ, we indicate domains of α where self- oscillations of complicated structure arise. An example of the filter (D3, m = 3, $\gamma = 4$) is represented in Fig. 1. The cross-hatched domain shows that there are oscillations of complicated structure for the corresponding values of α.

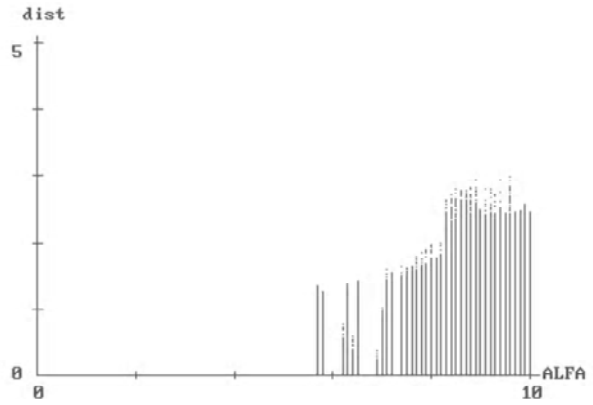


Figure 1. Evolutionary D3 ($m=3, \gamma=4$) shows that there are oscillations of complicated structure for $\alpha \geq 5.8$.

The aperiodic attractor of a chaotic type for $\alpha = 8.5$ is represented in Fig. 2.

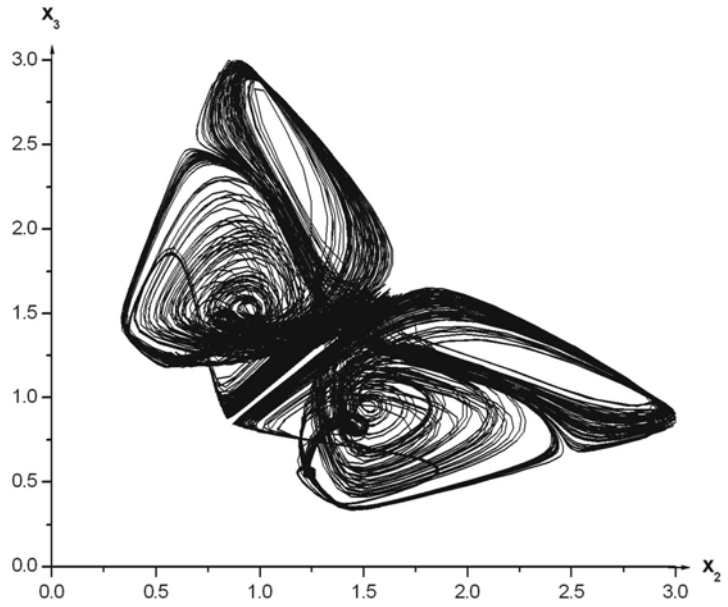


Figure 2. An aperiodic attractor of the system (2) for $m=3, \alpha=8.5, \gamma=4$.

DISCUSSION

Study of appearance conditions of chaos in deterministic systems is one of the fundamental problems of modern science. In the present paper we propose a mathematical model describing genetic systems consisting of three genetic elements. The model is a system of three ordinary differential equations. Unlike known autonomous systems with chaotic regimes of evolution, that system consists of completely identical (in form and parameters) equations and is symmetric with respect to permutation of variables. As a

function describing the regulatory mechanism we use the function proposed by Mackey and Glass in 1977 to simulate physiological processes. This function is ideal for description of genetic processes; in particular, processes of expression effectiveness regulation of genetic elements. The regulatory mechanism described by the function has the following characteristics: a small concentration of regulator gives expression activation, a large concentration – degradation. This mechanism is not exotic for living systems. For example, regulation of bacteriophage lambda promoter Prm efficacy has the same type (Ptashne *et al*., 1980). Therefore our model shows that construction of symmetric genetic makes from natural elements with behavior of chaotic types is fundamentally important. Note that, looking for necessary behavior, instead of using the very function we take its superposition of three order ($m = 3$), which complicates essentially the genetic mechanism because of its sensitivity to variations of concentrations of regulators. The result obtained in our paper poses the problem of study of relationships between symmetry and chaos in living (in particular, genetic) systems. We raise also the question about the role of chaos in functioning and evolution of living systems.

ACKNOWLEDGMENTS

The work was supported in part by the Russian Foundation for Basic Research (No. 04-01-00458), NSF:FIBR (Grant EF–0330786) and by the Russian Federal Agency for Science and Innovations (State contract No. 02.467.11.1005).

REFERENCES

Akhromeeva T.S., Kurdyumov S.P., Malinetskii G.G. (1988) Paradoxes of the world of nonstationary structures. In *Computers and Nonlinear Phenomena: Informatics and Modern Natural Sciences*, Nauka, Moscow, 44–122.

Fadeev S.I. Pokrovskaya S.A., Berezin A.Yu., Gainova I.A. (1998) *The Package STEP for Numerical Study of Systems of Nonlinear Equations and Autonomous System of General Form. Description of work of the package STEP by examples of problems from the course "Engineering chemistry of catalytic processes"*. Novosibirsk State University Publ., Novosibirsk.

Mackey M.C., Glass L. (1977) Oscillation and chaos in physiological control systems. *Science*, **197**, 287–289.

Ptashne M., Jeffrey A., Johnson A.D., Maurer R., Meyer B.J., Pabo C.O., Roberts T.M., Sauer R.T. (1980) How the lambda repressor and cro work. *Cell*, **19**, 1–11.

SEARCHING CONSTRAINTS IN BIOLOGICAL REGULATORY NETWORKS USING SYMBOLIC ANALYSIS

Mateus D.*[*]*, Gallois J.P.
CEA/LIST Saclay, F-91191 Gif sur Yvette Cedex, France
[*] Corresponding author: e-mail: daniel.mateus@cea.fr

Key words: regulatory networks, symbolic analysis, kinetic logic, temporal logic

SUMMARY

Motivation: Tools for modeling and simulation are needed to understand the functioning of biological regulatory networks. The difficulty of determining the parameters of the models motivates the use of automatic methods able to find the constraints on the parameters of a model that matches the behavior of the actual system.

Results: The method is applied on the qualitative modeling approach known as generalized kinetic logic. The logical parameters of the model, which are related to the kinetic parameters of a differential description, can be unknown. Translating the model into a symbolic transition system, and the known behaviors into temporal logic formulas, we can find the constraints on the logical parameters corresponding to all the logical models having the specified behavior. We illustrate the pertinence of the method on the published example of the mucus production of *Pseudonomas aeruginosa*.

Availability: AGATHA is the property of CEA; Contact the authors for further information.

INTRODUCTION

Modeling and simulation can be necessary to understand genetic regulatory networks as the complexity of the interleaved interactions between constituents of the network (as genes and proteins) makes intuitive reasoning too difficult. A way to check the model against the actual behavior is the simulation, which can be performed by computer tools (de Jong, 2002).

One typical bottleneck in the modeling approach is the difficulty to determine parameters which can change the simulated behavior (as kinetic parameters in differential equations). It appears interesting to use qualitative models, less precise than quantitative ones, but where the simulations remain interesting. In generalized kinetic logic (Thomas, D'Ari, 1990), the concentrations of the constituents are represented by variables which can only take a finite number of integer values. Such a logical model can be seen as a qualitative abstraction of a system of piecewise-linear differential equations (Snoussi, 1989). There is a tool dedicated to the simulation and analysis of these models, called GINsim (Chaouiya *et al.*, 2006; Larrinaga *et al.*, 2006).

One of the difficulties of the modelling task remains the determination of the logical parameters, upon which the evolution depends. Properties have been proved to be useful in this field (Thomas *et al.*, 1995), and are used in GINsim. An additional way to proceed is to exhaustively generate all the models (all the possible parameters values), and automatically choose the models having a specified behaviour (Bernot *et al.*, 2004). We

propose a complementary approach, permitting to restrict the possible values of the parameters given a specified behavior. An incomplete model, e.g. a model whose logical parameters are unknown, can be translated into a symbolic transition system (STS). The known constraints on parameters can be specified in this STS. Using symbolic execution techniques, all the behaviors permitted by the constraints can be generated. Moreover, each behavior, or *path*, is associated with a *path condition*, which is the constraint on parameters permitting this behavior. Adapting model-checking techniques (Clarke *et al.*, 1999), all the paths verifying a specified behavior (given by a linear temporal logic formula), are selected, and the associated path conditions are reduced into a single constraint. The parameters which verify the constraints are those leading to a model which has the specified behavior. The constraint can be added to the symbolic transition system associated with the regulatory network, and the resulting model can be simulated.

The method is illustrated with the network of the mucus production in Pseudonomas aeruginosa (Bernot *et al.*, 2004). This work has been implemented in the AGATHA tool, which is also used for the verification and test of industrial specifications (Bigot *et al.*, 2003).

MODELS

We present briefly the logical modeling of regulatory networks, which can be found in Thomas and D'Ari (1990). A logical description is constituted of n variables, each representing the concentration of a constituent of the actual network, mainly the proteins produced by the genes of the network. Each variable x_i can take an integer value between 0 and b_i. A logical state $E = (E_1, ..., E_n)$ is a vector of values of the variables. With each state E, and each variable x_i, is associated a logical parameter $K(x_i, E)$, which has an integer value between 0 and b_i. The logical parameter is the value toward which the associated variable tends in the associated logical state. It means that in the logical state E:

If $K(x_i, E) > E_i$, then $(E_1, ..., E_i + 1, \ldots, E_n)$ is a successor of E;

If $K(x_i, E) < E_i$, then $(E_1, ..., E_i - 1, \ldots, E_n)$ is a successor of E;

If $K(x_i, E) = E_i$, for all i, then E is called a steady state, and has no successor.

The graph of sequences of states is constituted of the logical states, and the transitions between each state and its successors.

As the logical description is related to a differential one, there are constraints on the parameters: some equalities and inequalities between logical parameters are known. In particular, when a graph of interactions is known, e.g. when positive or negative interactions between genes are known, some constraints can be deduced (the example below illustrates this point).

METHODS

Given a constraint C on the logical parameters, and an initial logical state E, we generate a symbolic transition system (STS). Then the symbolic execution of the STS is made. This method constructs a tree of sequences of logical states, with the following rules:

The root of the tree is the initial state E;

For each possible successor of E, there can be a path constructed, if and only if the condition D on the logical parameters that makes a logical state E' a successor of the initial state is compatible with C; then E' is constructed, and an edge is constructed from E to E';

E' is associated with a new constraint C', which is the conjunction of C and D;

The process is repeated with the successors of E' and the constraint C';

If a new logical state has already been reached in the same path, then the execution of this path stops;

The symbolic execution is over when all the possible paths have been treated.

We see that every state in the tree is associated with a constraint, which is called *path condition*, and is the constraint on the parameters which is necessary to the existence of the associated path in the logical model of the network.

To search a specific path in the symbolic execution tree we have adapted model-checking techniques for Linear Temporal Logic (LTL). A LTL formula expresses properties of a path. This logic adds to the classical operators of propositional logic mainly two temporal operators, called Next (N), and Until (U). If f and g are formulas, Nf means that f is true in the following state of the path, and fUg means that f is true in each state of the path, until g becomes true (and g eventually happens). We can then define the operators Finally (F) and Globally (G); Ff means that f eventually happens (and can be written $TrueUf$); Gf means that f is always true (and can be written $Not(F(Not\, f))$.

The method we use selects all the paths verifying the LTL formula, and synthesizes the disjunction of the path conditions associated with the last state of each path. The resulting constraint represents all the parameters compatible with the behaviour specified by the formula.

RESULTS

We illustrate the results on the example of the mucus production in *Pseudonomas aeruginosa* (Bernot *et al.*, 2004). These bacteria secrete mucus only in lungs affected by cystic fibrosis. The simple regulatory network associated with the mucus production contains the protein AlgU, and an inhibitor complex anti-AlgU. AlgU has a positive effect on anti-AlgU and on itself, while anti-AlgU has a negative effect on AlgU. It means that a high concentration of AlgU tends to increase the rates of transcription associated with anti-AlgU and AlgU itself, while a high concentration of anti-AlgU tends to decrease the rate of transcription associated with AlgU. A sufficient concentration of AlgU leads to the production of mucus.

The logical description of the associated network contains two variables x and y, respectively corresponding to AlgU and anti-AlgU. x can take the values 0, 1, 2, and y can take the values 0, 1. We assume that if $x > 0$ then x has a positive effect on y, if $x = 2$ then x has a positive effect on x, and if $y = 1$ then y has a negative effect on x. Moreover when $x = 2$, the production of mucus is possible. Fig. 1 summarizes these interactions.

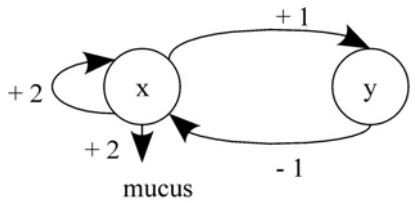


Figure 1. Graph of interactions.

These hypotheses are translated into constraints with the following rules: if between two states E and E', the interactions on x are the same, then $K(x, E) = K(x, E')$; if in E there is a negative interaction on x that is not in E', or if in E' there is a positive interaction on x that is not in E, then $K(x, E) \leq K(x, E')$; the same rules applies for y. For example, $K(x,(0,1))=K(x, (1,1))$ and $K(x, (1,1)) \leq K(x, (1,0)) \leq K(x, (2,0))$. Now we translate the expected behaviours in LTL. As bacteria do not produce mucus in a common environment, there is no path from a state where $x = 0$ to a state where $x = 2$. A path from $x = 0$ to $x = 2$ is described by the formula $((x = 0)$ and $F(x = 2))$. Moreover bacteria which produce mucus can continue to produce mucus even in a different environment

(e.g. isolated from infected lungs). A path beginning with $x = 2$ which turns back forever to $x = 2$ is described by $((x = 2)$ and $G(F(x = 2)))$.

The resulting constraint can be written $K(x,(0,0)) < 2$ and $((K(x, (2,1)) = 2$ and $K(y, (1,1) = 1)$ or $(K(x, (2,0)) = 2$ and $K(y, (1,1)) = 0))$. There are 8 different models verifying the constraint.

DISCUSSION

Our approach permits the simulation of a partially known logical description of a genetic regulatory network. Giving a hypothesis on the behaviour of the system, and translating it into a temporal logic formula, it is possible to determine the constraints on the unknown parameters associated with this behaviour. The method can be incremental as the new constraint can be added to the previous model. We present an example of a small network but we have worked on the model of the immunity control in bacteriophage lambda (a four genes model) (Thieffry, Thomas, 1995). We are working on refining some of the tools we use to be able to treat bigger networks.

ACKNOWLEDGEMENTS

The authors are grateful to D. Bahrami, N. Rapin, P. Le Gall and J.P. Comet for helpful discussions.

REFERENCES

Bernot G., Comet J.P., Richard A., Guespin J. (2004) Application of formal methods to biological regulatory networks: extending Thomas' asynchronous logical approach with temporal logic. *J. Theor. Biol.*, **229**, 339–347.

Bigot C., Faivre A., Gallois J.P., Lapitre A., Lugato D., Pierron J.Y., Rapin N. (2003) Automatic test generation with AGATHA. In *TACAS, LNCS*, **2619**, 591–596.

Chaouiya C., Thieffry D, Sánchez L. (2006) From gradients to stripes: a logical analysis of Drosophila segmentation genetic network. In Kolchanov N., Hofestaedt R., Milanesi L. (eds.), *Bioinformatics of Genome Regulation and Structure II*, Springer, 23–24.

Clarke E.M., Grumberg O., Peled D. (1999) *Model checking*, Cambridge, Mass., MIT Press.

de Jong H. (2002) Modelling and simulation of genetic regulatory systems: a literature review. *J. Comput. Biol.*, **9**, 67–103.

Larrinaga A., Naldi A., Sanchez L., Thieffry D., Chaouiya C. (2006) GINsim: A software suite for the qualitative modelling, simulation and analysis of regulatory networks. *Biosystems*.

Snoussi E.H. (1989) Qualitative dynamics of piecewise-linear differential equations: A discrete mapping approach. *Dyn. Stability Syst.*, **4**, 189–207.

Thieffry D., Thomas R. (1995) Dynamical behaviour of biological regulatory networks II. Immunity control in bacteriophage lambda. *Bull. Math. Biol.*, **57**, 277–97.

Thomas R., D'Ari R. (1990) *Biological feedback*, Boca Raton, CRC Press.

Thomas R., Thieffry D., Kaufman M. (1995) Dynamical behaviour of biological regulatory networks I. Biological role of feedback loops and practical use of the concept of the loop-characteristic state. *Bull. Math. Biol.*, **57**, 247–276.

MATRIX PROCESS MODELLING: DEPENDENCE OF SOLUTIONS OF A SYSTEM OF DIFFERENTIAL EQUATIONS ON PARAMETER

***Matveeva I.I.*[1], *Popov A.M.*[2]**
[1] Sobolev Institute of Mathematics, SB RAS, Novosibirsk, 630090, Russia; [2] Novosibirsk State University, Novosibirsk, 630090, Russia
[*] Corresponding author: e-mail: likho@bionet.nsc.ru

Key words: genetic systems, mathematical models, ordinary differential equations, Cauchy problem, delay differential equations

SUMMARY

Motivation: Matrix processes of replication, transcription, and translation are essential part of natural and artificial genetic systems. To reflect adequately these processes in models of gene networks it is necessary to describe hundreds and thousands elongation steps. 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 present paper we analyze a mathematical model describing a regulated process of matrix synthesis of RNA and proteins.

INTRODUCTION

Gene networks are complex multicomponent systems whose core is gene elements mutually regulated. One of the fundamental processes of synthesis of RNA and proteins encoded by gene elements are matrix elongation processes that are long unbranched chains of successive nonlinear processes of tailing of nucleotides and amino acid residues to growing chains of RNA and proteins, respectively. To reflect adequately matrix processes in models of gene networks constructed on the basis of chemical-kinetic representations in terms of differential equations it is necessary to use hundreds and thousands of additional equations. In this connection we have the important problem of development of more economic tools for description of elongation stages in models of gene networks. One of approaches is to replace systems of differential equations by one delay differential equation with a parameter arising as a delay in the model. This approach is based on the fact that, because of elongation stages, active forms of macromolecules of RNA and proteins arise only after a time from the beginning of the synthesis; moreover, under some conditions, the process time must be taken into account. As a rule, we disregard polymerization mechanism (complex enough in real systems). As was established, this intuitive assumption has the rigorous mathematical substantiation (Likhoshvai *et al.*, 2004). Consider the simplest model of gene expression

$$\begin{aligned}
&\frac{dx_1}{dt} = -\frac{n-1}{\tau}x_1 + g(x_n),\\
&\frac{dx_i}{dt} = \frac{n-1}{\tau}x_{i-1} - \frac{n-1}{\tau}x_i, \quad i = 2,...,n-1,\\
&\frac{dx_n}{dt} = \frac{n-1}{\tau}x_{n-1} - \theta x_n,
\end{aligned} \tag{1}$$

where $\tau > 0$, $\theta > 0$, the function $g(u)$ is bounded and satisfies the Lipschitz condition:

$$\sup_{u\in R}|g(u)| = G, \qquad |g(u_1) - g(u_2)| \le L|u_1 - u_2|.$$

For a fixed τ, the sequence of the last components $x_n(t,\tau)$ of solutions of a series of the Cauchy problems with the zero initial conditions

$$x_1|_{t=0} = ...x_n|_{t=0} = 0 \tag{2}$$

converges uniformly to a limit function $x(t,\tau)$

$$\lim_{n\to\infty} x_n(t,\tau) = x(t,\tau) \tag{3}$$

as the number n of equations in (1) increases unrestrictedly; moreover, the limit function $x(t,\tau)$ is a solution of the initial problem for the delay differential equation

$$\begin{aligned}
&\frac{dx(t,\tau)}{dt} \equiv -\theta x(t,\tau) + g(x(t-\tau,\tau)), \qquad t > \tau,\\
&x(t,\tau) = 0, \qquad t \in [0,\tau].
\end{aligned} \tag{4}$$

The convergence (3) in the paper (Likhoshvai *et al.*, 2004) was established on the interval $[0,T]$, where $T > \tau$ is such that

$$L\left(\frac{1-e^{-\theta T}}{\theta}\right) < 1. \tag{5}$$

The present paper continues to study (Likhoshvai *et al.*, 2004) solutions of the system of differential equations (1). We conduct a comparative analysis of two systems of the form (1) with the same number of equations with various values of the parameter τ. We estimate difference between the last components of solutions of the Cauchy problems for these systems on the interval [0,T].

RESULTS

In the present paper, for a fixed n, we study properties of the last component $x_n(t,\tau)$ of the solution of the Cauchy problem (1), (2) as a function of t and the parameter τ.

Theorem 1. *Let* τ_1, $\tau_2 > 0$ *be such that*

$\frac{\theta\tau_j}{n-1} < 1$, $j = 1, 2$.

Then the estimate holds

$\max\limits_{t\in[0,T]} \left|x_n(t,\tau_1) - x_n(t,\tau_2)\right| \le c\left|\tau_1 - \tau_2\right|$,

where $c > 0$ *is a constant independent of* τ_1, τ_2, $T > \max\{\tau_1, \tau_2\}$ *satisfies the inequality (5).*

By completeness of $C[0,T]$, From Theorem 1 the uniform convergence follows

$$\lim_{n\to\infty} x_n(t,\tau) = x_n(t), \quad 0 \le t \le T; \tag{6}$$

moreover,

$\max\limits_{t\in[0,T]} \left|x_n(t,\tau) - x_n(t)\right| = O(\tau)$, $\tau \to 0$.

One of natural questions arises: What can we say about differential properties of the limit function $x_n(t)$? The following theorem answers the question.

Theorem 2. *Let* $T > 0$ *satisfy the inequality (5). For the limit function* $x_n(t)$ *we have the identity*

$\frac{dx_n(t)}{dt} \equiv -\theta x_n(t) + g(x_n(t))$, $\quad 0 < t < T$,

and $x_n(0) = 0$.

Note that θ and the function $g(u)$ do not depend on n. Consequently, by Theorem 2, the function $x_n(t)$ does not depend on n, i. e.,

$$x_n(t) \equiv x(t), \tag{7}$$

where $x_n(t)$ is a solution of the Cauchy problem for an ordinary differential equation

$$\begin{aligned} &\frac{dx}{dt} \equiv -\theta x + g(x), \quad 0 < t < T, \\ &x\big|_{t=0} = 0. \end{aligned} \tag{8}$$

As a result we have the next questions: What can we say about behavior of the sequence of the functions $x(t,\tau)$ defined by the convergence (3) as $\tau \to 0$? Does the limit exist? Does the limit coincide with the function $x(t)$? What differential properties does the limit possess? We answer these questions below.

Theorem 3. *Let* $T > 0$ *satisfy the inequality (5). The sequence* $\{x(t,\tau)\}$ *converges uniformly on the interval* $[0,T]$ *as* $\tau \to 0$*:*

$$\lim_{\tau\to 0} x(t,\tau) = y(t); \tag{9}$$

moreover, the limit function $y(t)$ *is a solution of the Cauchy problem (8).*

In the present paper we establish an estimate for rate of the convergence (9).

Theorem 4. *We have the limit relation as* $\tau \to 0$

$$\max_{t\in[0,T]} \left| x(t,\tau) - y(t) \right| = O(\tau) .$$

By uniqueness of the Cauchy problem, we obtain that the function $x(t)$, defined by (7), and the function $y(t)$, defined by (9), coincide. Thus, from Theorem 1–3 we have the following result.

Theorem 5. *For the last component of the solution of the Cauchy problem (1), (2) we have the relation*

$$\lim_{\tau\to 0}\lim_{n\to\infty} x_n(t,\tau) = \lim_{n\to\infty}\lim_{\tau\to 0} x_n(t,\tau) .$$

DISCUSSION

Development of new methods for modeling and analysis of models describing dynamics of gene networks is one of actual problems of systemic biology. Within the framework of the subject there exists an important problem concerning estimating errors between various models describing the same processes. In the paper we have presented results of solving such a problem for a system modeling matrix processes in the cell. In future, we plan to generalize the result on the case when transition from the i th stage to the $i+1$st stage is defined by nonlinear functions of the form $\frac{n-1}{\tau} x_i f_i$, where f_i is a continuous nonnegative bounded below function of several variables (may be one variable x_i). At first we are going to study our problem for the function $f_i = \frac{1}{1+x^{\gamma}}$. Solving the problem will allow us to come to substantiation of the following important methodological approach: description of a successive chain of nonlinear processes where each elementary step is defined by its own nonlinear function can be conducted by means of systems of the form (2).

ACKNOWLEDGEMENTS

The work was supported by the Russian Foundation for Basic Research (Nos 05-04-49068, 05-07-90274), the Siberian Branch of the Russian Academy of Sciences (Interdisciplinary integration project No. 24). The authors are grateful to Genady Demidenko and to Vitaly Likhoshvai for helpful discussions.

REFERENCES

Likhoshvai V.A., Demidenko G.V., Fadeev S.I., Matushkin Yu.G. (2004) Modelling of multi-stage synthesis without branching by an equation with delay. *Sib. Zh. Ind. Mat.*, **7**, 73–94. (In Russ.).

AN INTEGRATION OF THE DESCRIPTIONS OF GENE NETWORKS AND THEIR MODELS PRESENTED IN SIGMOID (CELLERATOR) AND GENENET

***Podkolodny N.L.*[*1, 2], *Podkolodnaya N.N.*[1], *Miginsky D.S.*[1], *Poplavsky A.S.*[1], *Likhoshvai V.A.*[1], *Compani B.*[3], *Mjolsness E.*[3]**

[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia; [2] Institute of Computational Mathematics and Mathematical Geophysics, SB RAS, Novosibirsk, Russia; [3] Institute for Genomics and Bioinformatics, University of California, Irvine, USA

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

Key words: gene network, metabolic pathway, model, databases

SUMMARY

Motivation: The problems that arise when modeling complex molecular genetic systems at the cell level are so large-scale that they require integrated efforts of many research teams. Therefore, it is a topical problem to integrate the technologies applied to description of gene networks and their models developed at the ICG (Russia) and UCI (USA) with the aim to provide a tight cooperation in the field of systems biology.

Results: A software for conversion of gene networks and the models of gene network dynamics represented in the GeneNet system into a format suitable for loading to the SIGMOID database and Cellerator system was developed.

INTRODUCTION

An original software for supporting the technological chain of modeling was designed at the Institute of Cytology and Genetics, SB RAS, including creation of databases compiling the descriptions of various organization levels of genetic systems, formalization (specification) of the models of genetic systems, study of the models' behavior, search for the models' parameters according to the experimentally observed gene network behavior, accumulation of basic models, and solution of the target problems. A language for the specification, SiBML, was developed; it is oriented to the construction of mathematical models of molecular genetic systems taking into account the main specific features of their structure: a linear ordering and gene orientations within the genomes, polyvariance of genes (polyallelism), and polycompartment pattern of biological systems. An original software supporting SiBML for computing the direct problem was developed as well as for solving the problem of verification of mathematical models of gene networks. An original technology GeneNet (Ananko *et al*., 2002, 2005) that enables accumulation of data in a database with a consequent analysis of heterogeneous information on gene and metabolic networks was designed. A large number of the gene networks describing the vital molecular genetic processes were reconstructed and are presented in the GeneNet database.

A software for systems biology SIGMOID (Cheng *et al*., 2005; http://www.sigmoid.org/) that in turn calls Cellerator (Shapiro *et al*., 2003) was developed

at the University of California, Irvine; this software provides a wide range of options for description of biological processes and their mathematical models.

An integration of the gene networks and their models presented in GeneNet and SIGMOID will allow for an efficient combination of the technologies for modeling genetic systems developed at ICG and the technologies for distributed modeling that are developed at UCI.

THE REPRESENTATIONS OF GENE NETWORKS IN GENENET AND SIGMOID DATABASES

Functioning of a gene network is provided by complex relationships between different components, namely, genes, proteins, metabolites, signal molecules, energy-connected cell components, etc. Using an object-oriented approach, we recognize several following logical levels in the description of relationships between the gene network components. Ontological level, including general notions and relations between them. Here we describe as metaclasses the elementary structures, or Entities (genes, proteins and protein complexes, RNAs, and small molecules) and the elementary processes (reactions and regulatory events). A scheme of semantic relationships between the elementary structures and processes is given in Fig. 1.

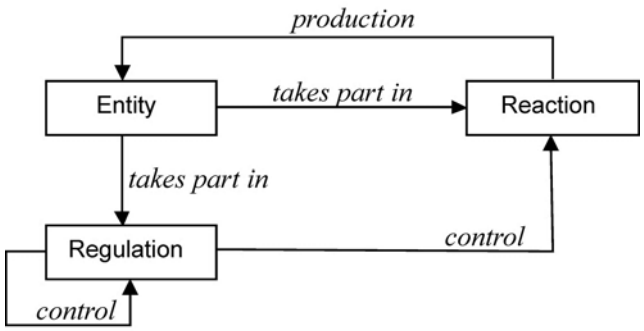


Figure 1. Semantic relations between elementary notions in a gene network.

The level of objects of study involves the descriptions of notions and classes of entities for particular objects of study. For example, description of some representatives of such classes as “genes”, “proteins and protein complexes”, “RNAs”, and “small molecules”, which are involved in functioning of particular objects studied (i.e., cells of *E. coli* K12).

The level of compartments, at which the entities described above are attached to a spatial compartment. At this level, the description of entities may include supplementary parameters such as, for example, the concentration of a given entity in a particular compartment. The whole bulk of information about elementary structures and functional relations in a gene network is represented at the three levels described above. The subsequent two levels are needed for describing the network as a whole.

The level of gene networks corresponds to the description of functional subsystems. At this level, the particular descriptions of a complex system (the object of study) from different viewpoints may be represented as well as its simplified description or a partial model representation. Many representations of a complex system may exist. Besides, there is a possibility of the gene network reconstruction via a query to the GeneNet database.

The level of representation provides the way for describing the pattern of visualization or graphical representation of a gene network. It is supposed that different ways of representation exist for each gene network. In particular, to this level we refer the representation of a gene network in a form of a tree or hierarchical relations, in a form of a hyper graph, and so on. At this level, it is possible to use different ways of layouts for the automated visualization of a gene network, obtained by querying the database.

GeneNet format is based on XML. Each separate XML file represents a single GeneNet diagram and consists of the following sections:

Header (<diagram> element). It includes the identifier of a diagram in GeneNet database, dates of its creation and modification, and the description of biological function of the gene network.

Nodes list (<objects> element). This section contains the descriptions of all nodes in graph representing diagram. Nodes are genes, proteins, substances, RNAs, as well as reactions and any regulatory events (reaction or regulation). Each description includes name and synonyms, links to articles, comments, and some service information. Moreover, the reactions and regulatory events contain information about their inputs and outputs. The reactions in an object representation of the SIGMOID (see the scheme of information representation in the Sigmoid database at are divided into two subclasses—biological and mathematical reactions. The subclass "Biological reactions" is designed for the description of various reaction types with a distinct biological interpretation (for example, replication, transcription, translation, allosteric interaction, enzymatic reaction, etc.). A mathematical representation of the model described in the subclass "Mathematical reactions", containing the hierarchy of methods for modeling biological reactions, is available to mathematically model each of the "Biological reaction" subclasses.

SiBML, A LANGUAGE FOR SPECIFICATION OF MODELS

The language SiBML is designed for an efficient (economic) specification of the models; it comprises three following description levels: (1) the level for description of elementary processes; (2) the level for description of "genetic maps" (G maps), which are constructed as the lists of ordered oriented objects named "genes", each "gene" carrying the information about the elementary processes related to it; and (3) the level of specification of a compartment structure of the object modeled. At the last level, a map of intercommunicating compartments (or C map) is specified. The C map is constructed as an ordered list of objects named "compartments". Each "compartment" as an object contains the information about its own name, the compartments whereto the substances flow from this compartment, the information about G maps localized to this compartment, and the information about the files where the description of elementary processes are stored. The corresponding descriptions are stored in the databases at each level and can be used multiply. The models are assembled by a specialized set of programs, named model constructor. The models that appear when describing the processes in terms of chemical kinetic reactions are formally belong to the class of autonomous systems of differential equations. However, in a general case, the models belong to a mixed type, since they may contain continuous, discrete, probabilistic, and other modules. The final constructed model is an ordered list of elementary processes. It is this pattern that is convertible into the CELLERATOR format without any losses.

IMPLEMENTATION AND RESULTS

1. A software providing the loading of the gene networks contained in the GeneNet system into the SIGMOID database was developed. This brings about the problems of matching the objects described in these databases and the attributes of these objects,

Object. The <gene>, <rna>, <protein>, and <substance> elements represented in GeneNet are mapped into SIGMOID as interfaces Gene, RNA, Protein, and Molecule, respectively. Using SIGMOID API, it is possible to map name, species, references to papers, and comments.

Reaction. This class of regulatory events is represented in GeneNet as <reaction> elements; it is mapped in SIGMOID using the CatalyticWithAllostericRegulation interface. However, the following data can be mapped directly: substrates, products, references to papers, and comments. Additional information about enzymes, activators, and inhibitors, necessary for specification of CatalyticWithAllostericRegulation interface is extracted from GeneNet through the analysis of regulatory relationships (see below).

Gene regulation. In GeneNet, this is represented by the <reaction> element, whose output (i.e., the object of regulation) is the reaction of transcription, translation, etc. or the indirect reactions that include all the stages involved in gene expression. The reaction of transcription has an input (single gene), output (RNAs), and a set of transcription factors (proteins or protein complexes). This reaction can be mapped to SIGMOID as the RegulatoryRelationship interface with the following fields: target gene, regulators, references to papers, and comments.

Regulatory event. In GeneNet, this is a special type of interaction (<reaction> element) where the inputs are objects and the output is another interaction. Regulatory events may be positive or negative as well as direct or indirect (which means the lack of precise information about the particular mechanism). Regulatory event could be organized in complex multilevel cascades (regulation of regulation of …). In SIGMOID regulatory elements participate in reactions as additional inputs. They are not consumed and therefore they exit the reaction as outputs. Since GeneNet and SIGMOID possess different perspectives on how regulation should be modeled, GeneNet regulatory events are mapped in SIGMOID as activators, inhibitors, enzymes, or regulators of reactions.

2. A converter was designed able to convert the mathematical models constructed using a limited version of the SiBML standard into the standard of CELLERATOR software package aiming to further loading into the package Mathematica 5.0.

The input format of the converter is specified by the SiBML standard (Likhoshvai *et al*., 2001). Three files are used as the input data; these files contain the parameters (constants) and their values, dynamic variables and their initial values, and the corresponding mathematical model in SiBML, respectively.

The output of the converter operation is the file in a Cellerator standard organized as a NoteBook for Mathematica 5.0. The output NoteBook consists of a unit switching of the cellerator.m module, the unit containing the list of initial concentrations, and the unit with the list of reactions. The presented version of converter supports the following set of SiBML blocks: B(1->1); B(1=>1); B(0->1); B(0->1) ; B(UNI1) (see Table 1).

Table 1. Interpretation of standard SiBML blocks in terms of Cellerator

Block	Interpretation in Cellerator
B(UNI1)	$\left\{\left\{ 0 \to \mathrm{A_i},\quad -\mathrm{a_i} * \mathrm{G}\right\},\quad \left\{ 0 \to \mathrm{B_j},\quad \mathrm{b_j} * \mathrm{G}\right\}\right\},\quad \mathrm{i}=1,...,\mathrm{m};\ \mathrm{j}=1,...,\mathrm{l}$
B(1->1)	$\{A \to B,\quad K\}$
B(1=>1)	$\{0 \xrightarrow{B} A.\quad K\}$
B(0->1)	$\{0 \to A,\quad K\}$
B(1->0)	$\{A \to 0,\quad K\}$

ACKNOWLEDGEMENTS

The work was supported by the grant of NSF "FIBR: developmental modeling and informatics" and innovation project of the Federal Agency for Science and Innovation IT-CP.5/001 "Development of software for computer modeling and design in postgenomic system biology (system biology *in silico*)".

REFERENCES

Ananko E.A. *et al*. (2002) GeneNet: a database on structure and functional organisation of gene networks. *Nucl. Acids Res*., **30**, 398–401.

Ananko E.A. *et al*. (2005) GeneNet in 2005. *Nucl. Acids Res*., **33**, Database issue D425–D427.

Cheng J. *et al*. (2005) Sigmoid: A software infrastructure for pathway bioinformatics and systems biology. *IEEE Intelligent Systems*, **20**(3), 68–75.

Likhoshvai V.A. *et al*. (2001) A generalized chemical-kinetic method for modeling gene networks. *Mol. Biol. (Mosk.)*, **35**(6), 1072–1079.

Shapiro B.E. *et al*. (2003) Cellerator: extending a computer algebra system to include biochemical arrows for signal transduction simulations. *Bioinformatics*, **19**(5), 677–678.

SPECTRAL ANALYSIS OF GENE EXPRESSION PROFILES USING GENE NETWORKS

***Rapaport F.*[1, 2]*, ZinovyevA.*[*1, 3]*, Barillot E.*[1]*, Vert J.-P.*[2]**
[1] Institute Curie, Bioinformatics Service, Paris, France; [2] Ecole des Mines de Paris, Center for Computational Biology, Fontainebleau, France ; [3] Institute of Computational Modelling, SB RAS, Krasnoyarsk, Russia
[*] Corresponding author: e-mail: andrei.zinovyev@curie.fr

Key words: genetic networks; microarray analysis; spectral analysis; classification problem

SUMMARY

Motivation: Microarrays have become extremely useful for analysing genetic phenomena, but establishing a relation between microarray analysis results (typically a list of genes) and their biological significance is often difficult. Currently, the standard approach is to map *a posteriori* the results onto gene networks to elucidate the functions perturbed at the level of pathways. However, integrating *a priori* knowledge of the gene networks could help in the statistical analysis of gene expression data and in their biological interpretation.

Results: We propose a method to integrate *a priori* the knowledge of a gene network in the analysis of gene expression data. The approach is based on the spectral decomposition of gene expression profiles with respect to the eigenfunctions of the gene interaction graph. We applied the method to the analysis of a set of expression profiles from irradiated and non-irradiated yeast strains. It performed at least as well as the usual classification but provides much more biologically relevant results and allows a direct biological interpretation.

Availability: http://bioinfo.curie.fr/projects/kernelchip

INTRODUCTION

During the last decade microarrays have become the technology of choice for dissecting the genes responsible for a phenotype. One of the aims of microarray data analysis is to discover co-operating genes that are similarly affected during the experiment. Many databases and tools help verify this *a posteriori*, allowing a list of genes (for example, differentially expressed genes) to be crossed with gene networks, including metabolic, signaling or other regulation pathways. However, introducing biologically relevant information at this point in the analysis sacrifices some statistical power to the simplicity of the approach. For example, it is clear that a small but coherent difference in the expression of all the genes in a pathway should be more significant than larger but random differences.

There is a need for methods integrating *a priori* pathway knowledge in the gene expression analysis, and some attempts have been carried out in that direction so far. Pathway scoring methods are reviewed in (Curtis *et al*., 2005) and other examples of integrating pathway knowledge can be found in (Vert *et al*., 2003), (Hanisch, 2002).

In this paper we investigate a possibility for integrating genetic networks early in the gene expression analysis. We propose a method for calculating the eigen modes of the response of the gene network to a perturbation and suggest how these can be introduced

into supervised and unsupervised microarray data analysis. We illustrate the relevance of our approach by analysing gene expression data about transcriptional response of yeast colonies after low-dose irradiation. We obtain accurate and interpretable discriminative model such that the discriminative function behaves smoothly with respect to the graph of a genetic network. Here we give a short description of the methods and the results. For more detailed description, see (Rapaport *et al.*, 2006).

METHODS AND DATA

We consider a finite set of genes V of cardinality $|V| = n$. The available gene network is represented by an undirected graph $G = (V,E)$, in which the set of vertices V is the set of genes and $E \subset V \times V$ is the list of edges. Gene expression profiling gives a value of expression $f(u)$ for each gene u.

The Laplacian of the graph G is the $n \times n$ matrix:

$$L(u,v) = \begin{cases} number\ of\ neighbours\ of\ u, & if\ u = v \\ -1, & if\ (u,v) \in E \\ 0, & otherwise \end{cases}$$

We denote the eigenvalues of L by $0 = \lambda_1 \leq \ldots \leq \lambda_n$ and the corresponding eigenvectors by $e_1, \ldots, e_n$. The eigen-basis of L forms a Fourier basis. The Fourier transform $\hat{f} \in R^n$ of any expression profile f is defined by:

$$\hat{f}_i = \sum_{u \in V} e_i(u) f(u) \cdot \tag{1}$$

The discrete Fourier transform can be used for smoothing or for extracting features. We assume that the signal captured in the low-frequency component of the expression profiles contains the most biologically relevant information whereas the high-frequency components are more likely measurement noise. Let us use the following transformation:

$$S_\phi(f) = \sum_{i=1}^{n} \hat{f}_i \phi(\lambda_i) e_i \tag{2}$$

where $\phi : R^+ \to R$ is a non-increasing function that quantifies how each frequency is attenuated. If we put $\phi(\lambda) = 1$ for $\lambda \in [0; \lambda_0]$ and $\phi(\lambda) = 0$ otherwise, then we produce a low-pass filter that removes all frequencies from f above the threshold λ_0. Another forms of ϕ are also possible (Rapaport *et al.*, 2006).

In our study we use linear predictive models to predict a variable of interest y from an expression profile f that are obtained by solving the following optimisation problem:

$$\min_{\mathbf{w} \in \mathbf{R}^n} \sum_{i=1}^{p} l(\mathbf{w}^T \mathbf{f}_i, y_i) + C\|\mathbf{w}\|^2, \tag{3}$$

where $(\mathbf{f}_1,y_1),\ldots,(\mathbf{f}_p,y_p)$ is a training set of profiles containing the variable y to be predicted, and l is a quadratic loss function. The popular support vector machine and ridge regression are particular cases of this equation.

Here, we do not apply algorithms of the form (3) directly to the expression profiles f, but to their images $S_\phi(f)$. That is, we consider the problem:

$$\min_{\mathrm{w}\in\mathrm{R}^n} \sum_{i=1}^{p} l(\mathrm{w}^T S_\phi(\mathrm{f}_i), y_i) + C\|\mathrm{w}\|^2 \quad , \qquad \textit{which is equivalent to}$$

$$\min_{\mathrm{v}\in\mathrm{R}^n} \sum_{i=1}^{p} l(\mathrm{v}^T \mathrm{f}_i, y_i) + C \sum_{i:\phi(\lambda_i)>0} \hat{v}_i^2 \Big/ \phi(\lambda_i)^2 \,, \tag{4}$$

where $\mathrm{v} = \left(\sum_{i=1}^{n} \phi(\lambda_i)^2 \mathrm{e}_i \mathrm{e}_j^T\right)^{1/2} \mathrm{w}$ and $\hat{v}_i$ are the Fourier components of v. Thus, the resulting algorithm amounts to finding a linear predictor v that minimizes the loss function of interest *l* regularised by a term that penalises the high-frequency components of v. This is different to the classical regularisation $\|\mathrm{v}\|^2$ that only focuses on the norm of v. As a result, the linear predictor v can be made smoother on the gene network by increasing the parameter *C*. This allows the prior knowledge to be direcly included because genes in similar pathways would be expected to contribute similarly to the predictive model.

For testing the algorithm we collected the expression data from a published study analysing the effect of low irradiation doses on *S. cerevisiae* strains (Mercier *et al*., 2004). The first group of extracted expression profiles was a set of twelve independent yeast cultures grown without radiation, the second group was a set of six independent irradiated cultures exposed to a dose of 15–20 mGy/h for 20h. This dose produces no mutagenic effects, but induces transcriptional changes.

The metabolic gene network is a graph in which the enzymes are vertices and the edges between two enzymes indicate that the product of a reaction catalysed by the first enzyme is the substrate of the reaction catalysed by the second enzyme. We reconstructed this network from the KGML v0.3 version of KEGG (http://www.genome.jp/kegg/), resulting in 4694 edges between 737 genes. We kept only the largest connected component (containing 713 genes) for further spectral analysis.

RESULTS

We tested the performance of supervised classification after the transformation (2) with a support vector machine (SVM) trained to discriminate irradiated samples from non-irradiated samples. For each percentage of eigenvalues filtered, we estimated the performance of the SVM from the total number of misclassifications and the total hinge loss using a "leave-one-out" (LOO) approach (Vapnik, 1998). This approach removes each sample in turn, trains a classifier on the remaining samples and then tests the resulting classifier on the removed sample. For each fold, the regularization parameter was selected from the training set only by minimising the classification error estimated with an internal LOO experiment.

Fig. 1 (center) shows the classification results for the low pass eigenvectors filtering. We found that the performance of the classifier remained as accurate as the baseline performance until up to 80 % of the eigenvectors were discarded, with the hinge loss even exhibiting a slight minimum in this region. Overall these results show that classification accuracy can be kept high even when the classifier is constrained to exhibit coherence with the graph structure. On Fig. 1 we demonstrate that the resulting classifier behaves much more smoothly on the gene interaction graph, while having the same performance as the classical SVM. The biological interpretation of the classifier can be found in (Rapaport *et al*., 2006).

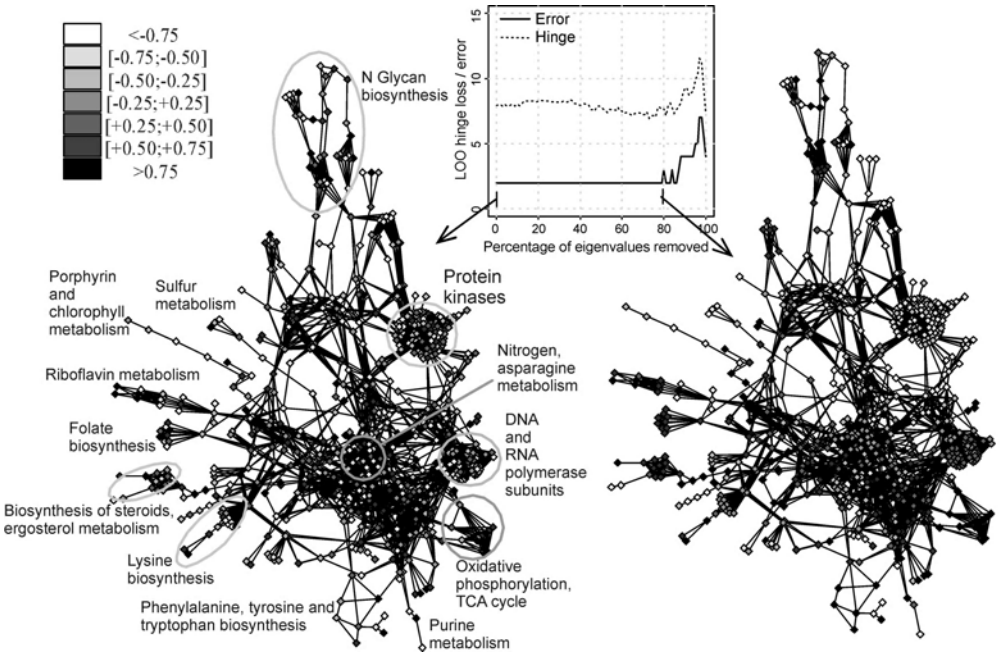


Figure 1. Global connection map of KEGG with mapped coefficients of the decision function v (equation 4) obtained by applying a customary linear SVM (left) and using high-frequency eigenvalue attenuation (80 % of high-frequency eigenvalues have been removed) (right). The graph of the misclassification error together with hinge function is shown in the center. The modified classifier behaves smoothly on the graph, with the activation of low-frequency eigen modes being determined by microarray data. Some parts of the network are annotated including big highly connected clusters corresponding to protein kinases and DNA and RNA polymerase sub-units. The upper-left part of the figure explains the color schema used to visualize the coefficient values.

DISCUSSION

Our algorithm groups predictor variables according to highly connected "modules" of the global gene network. We assume that the genes within a tightly connected network module are likely to contribute similarly to the prediction function because of the interactions between the genes. This motivates the filtering of gene expression profile to remove the noisy high-frequency modes of the network. This allows classifications based on functions, pathways and network modules rather than on individual genes, taking into account also pathways cross-talk. This leads to a more robust behavior of the classifier in independent tests and to equal if not better classification results.

ACKNOWLEDGEMENTS

This work was supported by the grant ACI-IMPBIO-2004-47 of the French Ministry for Research and New Technologies. We thank Sabrina Carpentier from the Service de Bioinformatique of the Institut Curie for the help she provided with the normalisation of the microarray data. We are very thankful for Dr. Marie Dutreix (Institute Curie, France) for providing data and fruitful discussions.

REFERENCES

Curtis K.R., Oresic M., Vidal-Puig A. (2005) Pathways to the analysis of microarray data. *Trends Biotechnol.*, **23**(8), 429–435.

Hanisch D., Zien A., Zimmer R., Lengauer T. (2002) Co-clustering of biological networks and gene expression data. *Bioinformatics*, **18**, Suppl. 1, S145–154.

Mercier G. *et al.* (2004) Biological detection of low radiation doses by combining results of two microarray analysis methods. *Nucl. Acids Res.*, **32**(1), e12.

Rapaport F., Zinovyev A., Dutreix M., Barillot E., Vert J.P. (2006) Spectral analysis of gene expression profiles using gene networks. *arXiv preprint q-bio.QM/0603030* (http://xxx.lanl.gov/abs/q-bio.QM/0603030).

Vapnik V. (1998) *The Nature of Statistical Learning Theory*. Springer-Verlag, New York.

Vert J.P., Kanehisa M. (2003) Extracting active pathways from gene expression data. *Bioinformatics*, **19,** Suppl. 2, II238–II244.

BIOPATH – A NEW APPROACH TO FORMALIZED DESCRIPTION AND SIMULATION OF BIOLOGICAL SYSTEMS

***Kolpakov F.*[*1, 2], *Sharipov R.*[1, 2, 3], *Cheremushkina E.*[2], *Kalashnikova E.*[3]**
[1] Institute of Systems Biology OOO, Novosibirsk, Russia; [2] Design Technological Institute of Digital Techniques, SB RAS, Novosibirsk, Russia; [3] Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia
[*] Corresponding author: e-mail: fedor@biouml.org

Key words: biological pathways, BioUML, cell cycle, NF-κB, arterial hypertension

SUMMARY

Motivation: Reconstruction of complex biological systems from a huge amount of experimental data requires a formal approach for description and simulation of biological pathways on different logical levels.

Results: Biopath database was developed for formalized description and simulation of biological systems using two new technologies BioUML and BeanExplorer. All Biopath data can be divided into 4 major blocks: regulation of eukaryotic cell cycle and cancer, NF-κB pathway and inflammation, nucleosomal regulation of gene expression, and human arterial hypertension. Biopath contains links to related biological databases and literature references. It is integrated with Cyclonet database – a specialized database on cell cycle regulation.

Availability: http://biopath.biouml.org.

INTRODUCTION

The main goals of the Biopath database are:

- describe formally structure and functioning of complex biological processes on different logical levels as a set of complementary diagrams;
- provide formal description of main properties of biological components (genes, proteins, substances, chemical reactions) that are components of these processes;
- simulate behavior of these processes or their parts to validate that we understand main principles of their functioning;
- provide user interface for Biopath access and editing through the Internet.

Currently we concentrate our attention on 4 biological processes: regulation of eukaryotic cell cycle and cancer, NF-κB pathway and inflammation, nucleosomal regulation of gene expression, and human arterial hypertension.

METHODS

Novel software technologies were used for development of Biopath database to achieve goals described above (Fig. 1):

- BioUML technology (Kolpakov, 2004; http://www.biouml.org) was used for formalized description of biological systems structure and functioning, visual modeling and editing of the database content. BioUML workbench also allows simulation the described systems behavior using Java or MATLAB simulation engines;
- BeanExplorer Enterprise Edition (http://www.beanexplorer.com) was used for development of web interface that provides user access to Biopath database trough the Internet (Fig. 2).

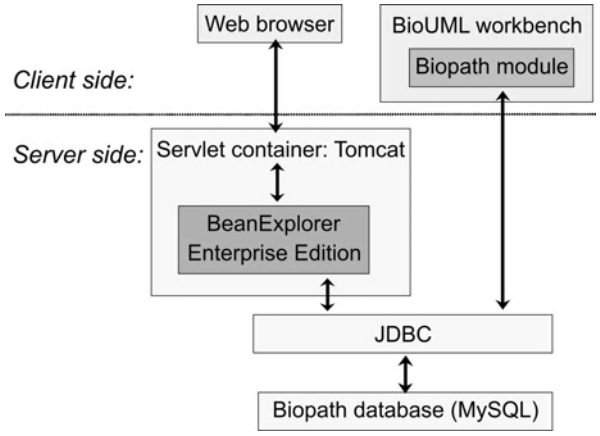


Figure 1. Biopath system architecture.

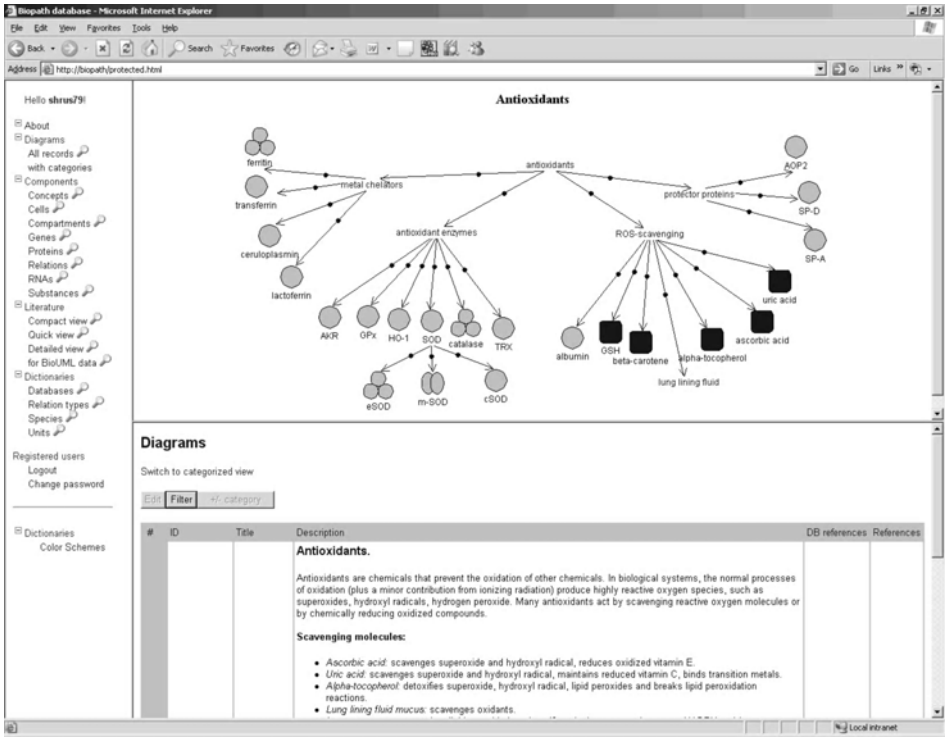


Figure 2. Web interface of Biopath database is generated with use of BeanExplorer EE technology.

Data collected in Biopath were compiled mainly from literature annotation. Genes, proteins and other entries in Biopath were supplied with links to the public databases GeneOntology, RefSeq and Ensembl. Cell cycle models stored in Biopath were imported

from SBML (Hucka *et al*., 2003; http://www.sbml.org) and CellML (http://www.cellml.org) model repositories or annotated from literature.

All data are stored in relational database (MySQL). Special database module was used to integrate Biopath database into BioUML workbench. Components of biological systems (mainly biological pathways) are classified similar to GeneNet database (Kolpakov *et al*., 1998) on following types: cell, compartment, proteins or protein complex, gene, RNA, substance and chemical reaction. Objects of each type are stored in separate table (Table 1). Concept and relation data types are used for description of semantic relationships between components of biological systems (Fig. 2). Diagrams and models are stored in diagrams table in DML format – specialized XML format used by BioUML workbench (Kolpakov, 2004). Additionally BioUML workbench generates image in PNG format and image map that are stored in diagrams table and used by BeanExplorer for generating web interface.

Table 1. Statistical report about Biopath database content

Table name and its brief description	Number of entries
Concepts - different biological concepts related with cell cycle regulations (e.g., G1 phase, mitosis, DNA replication)	1287
Cells - brief description of cell types and cell lines used in various experiments documented in Biopath	58
Compartments - brief description of cellular (e.g., nucleus, cytoplasm) and organism compartments (e.g., liver, blood)	156
Genes - quite comprehensive catalogue of genes involved in regulation of cell cycle or regulated in relation to cell cycle. The table includes also a manually annotated description of the role of the gene in cell cycle and other information	196
Proteins – description of proteins and their complexes involved in cell cycle. The table of activation or inhibitory role in the regulatory and signal transduction reactions	2452
Reactions - a catalogue of reactions between the biological components in the diagrams and models	2906
Relations - a catalogue of relations between the biological components in the diagrams and models	9337
RNA – description of various RNAs involved in cell cycle	30
Substances – description of substances used in experiments documented database	780
Diagrams (without models) - a catalogue of non-modeling diagrams	327
Models - a catalogue of diagrams modeling cell cycle and other biological processes	33
Publication references - a catalogue of articles used for diagram and their description creation	1341
References to other databases - a list of references to other international databases containing useful information about processes highlighted in Biopath	1850

To describe structure and functioning of complex biological processes on different logical levels and from different view points we are using following diagram types:

- semantic network (ontology) – describes semantic relationships between system components, system states and related problem domain concepts;
- pathway structure – describes structure of biological pathway as a compartmentalized graph. To facilitate understanding by human the pathway can be described as a set of complimentary diagrams where each diagram concentrates on some part of pathway structure or peculiarities of its functioning in specific cell types or conditions;
- pathway simulation – is extension of pathway structure diagram where different mathematical elements are associated with graph elements. This diagram type is used for automated code generation to simulate model behaviour by BioUML workbench.

RESULTS

Now Biopath database contains:

- 65 diagrams and 278 articles about NF-κB pathway, its regulation and participation in cellular processes and development of certain diseases;
- 56 diagrams and 434 articles about histones, "histone code", nucleosomal organization, DNA compaction and regulation of gene expression through histone modifications;
- 216 diagrams and 629 articles about regulation of eukaryotic cell cycle in normal and pathological states, including many types of cancer;
- 14 diagrams describing tyrosine hydroxylase pathway implicated in arterial hypertension development.

Statistics for main tables of Biopath database is shown in Table 1.

DISCUSSION

Suggested approach makes Biopath database suitable both for formal description of biological systems and their components like traditional databases on biological pathways (for example, KEGG, BIND and GeneNet) and for visual modelling of biological systems – majority of models that can be expressed on SBML or CellML can be mapped into corresponding Biopath diagrams. Their behaviour can be simulated using BioUML workbench and simulation results can be saved to Biopath database.

BioUML methodology allows a database annotator to use different diagram types to describe formally structure and functioning of complex biological processes on different logical levels as a set of complementary diagrams. Each diagram is provided with detailed description in HTML format and concentrates on some part of biological system or peculiarities of its functioning in specific conditions. Such approach greatly facilitates understanding of complex biological systems and processes by human.

Biopath has two user interfaces for the database access: through conventional web browser and through BioUML workbench. Using web browser user can query and edit the database content as well as view diagrams. BioUML workbench is used to query and edit the database content, view and edit diagrams, analyze and simulate the described systems behavior using Java or MATLAB simulation engines.

Traditionally SRS system (Erzold *et al*., 1996) is used for biological databases web publishing. BeanExplorer Enterprise Edition can be applied for the same task and provides following advantages in comparison with SRS system:

- BeanExplorer allows user to edit the database content through web interface;
- different operations (edit or delete selected records, data import and export, etc.) can be associated with different views;
- security and multi-user setup – user groups can have different privileges and roles in the system; different views and operations are available for different roles;
- hierarchical data classification – e.g., genes can be classified by their effect;
- seamless integration with relational databases. This allows integrating of several relational databases into one integrated system. An example of such integration is Cyclonet database where data on cell cycle and cancer from Biopath database are integrated with microarray and chemoinformatics data (Kolpakov *et al.*, 2006).

ACKNOWLEDGEMENTS

This work was supported by INTAS grant No. 03-51-5218, RFBR grant No. 04-04-49826-a and Siberian Branch of Russian Academy of Sciences (interdisciplinary project No. 46).

REFERENCES

Erzold T. *et al.* (1996) SRS: information retrieval system for molecular biology data banks. *Methods Enzymol.*, **266**, 114–128.

Hucka M. *et al.* (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, **19**, 524–531.

Kolpakov F.A. *et al*., (1998) GeneNet: a gene network database and its automated visualization. *Bioinformatics,* **14**, 529–537.

Kolpakov F.A. (2004) BioUML – open source extensible workbench for systems biology. *Proceedings of BGRS'2004*, **2**, 77–80.

Kolpakov F. *et al.* (2006) Cyclonet – an integrated database on cell cycle regulation and carcinogenesis. *Proceedings of the Fifth International Conference on Bioinformatics of Genome Regulation and Structure. This issue*.

RESEARCH ON BEHAVIOR OF GOVERNING GENE/EPIGENE NETWORKS AS A PROBLEM OF CELLULAR AUTOMATA IDENTIFICATION

Tchuraev R.N.*
Department of Physicochemical Biology and Epigenetics, Ufa Research Center, RAS, Ufa, Russia
* Corresponding author: e-mail: tchuraev@anrb.ru

Key words: governing gene networks; cellular automata and cell ensembles; metastability

SUMMARY

Motivation: A body of mathematics used in automata and graphs theories is adequate for revealing the general dynamic properties of governing gene and epigene networks and provides a basis for efficient analytical algorithms.

Results: The paper presents the results of research on the general properties of cellular automata characteristic functions as models for intracellular networks that govern gene expression.

INTRODUCTION

By now mathematical methods of different efficiency and purpose have been worked out for formalized description of gene network dynamics allowing the models be constructed for a number of real eukaryotic gene networks (Tchuraev, Galimzyanov, 2001; Ratushny *et al.*, 2003). Previously, as applied to intracellular governing gene networks of multicellular organisms, a simple formal system has been put forward, with cell governing networks represented as the inner structures of cellular automata A_e^c, and cell aggregates, being formed during ontogenesis, as the ensembles of cellular automata A_x. In the context of this theoretical model, we have stated the general principles of organization and dynamics laws of functioning in governing gene networks on the whole (Tchuraev, 2006).

For the purpose of investigating more thoroughly into the behavior of governing gene networks, the present paper deals with the general properties of a cellular automaton as a part of self-reproduced cell ensemble corresponded to a multicellular organism in the process of its individual development. These properties reveal the molecular and genetic mechanisms of storing, passing and transforming hereditary information during ontogenesis and phylogenesis.

MODEL

In order to reveal the general properties of cellular automata, the cellular automaton A_e^c, being an element of the cellular ensemble A_x, will be treated separately from specific elements of this ensemble. Then cells of a multicellular organism cultured *in vitro* may be taken as prototypes of such automata.

Let us introduce the premises necessary in constructing the models.

1. Let us assume that cells cultured *in vitro* and all their clonal derivatives formed under successive mitotic divisions are provided with substance and energy in amounts sufficient for reproduction, i.e., we believe that *resources of the outer medium impose no restrictions on intracellular metabolism (**postulate A**)*.

2. Let us also suppose that *all external effects performed by an experimenter are only signal ones (**postulate B**)*.

By doing so, any fluctuations of the nutrient medium composition are excluded, along with other external stochastic effects. External (input) signals may be carried by specific proteins and metabolites and/or temperature and other pulses.

3. Let us introduce one more rather evident premise (***postulate C***).

Every cytotype (molecular phenotype) in any cell of an organism x evolving in a neutral environment is determined by gene differentiated activity of a given genome.

Specific differentiated activity of the genome may be expressed, at each instant of time, by the γ-vector value of the activities $\mathbf{\Gamma}(t) = \langle \gamma_1(t), \ldots, \gamma_j(t), \ldots, \gamma_h(t) \rangle$ of genetic blocks G_j, elements of the intracellular governing gene network $S_e^g(G)$, which are considered to be the internal structure of cellular automaton $A_e^c \in A_x$. The governing gene network $S_e^g(G)$ may be transformed into the epigene network $\tilde{S}_e^e(G)$ that involves the following modules: genetic triggers (bistable memory modules), oscillators and delay logical combinators (Tchuraev, 2006).

A canonical description has already been given for the cellular automaton A_e^c, whose internal structure is represented by the governing gene/epigene network S_e^c.

In the general form the cellular automaton A_e^c is described with five symbols ($\mathbf{E}, \nabla, \Omega, \Phi, \Psi$), where $\mathbf{E}$ 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_e^c. *The input alphabet* $\mathbf{E}$ of the automaton A_e^c with n_1 number of input channels constitutes a set of corteges (words of length n_1): $\mathbf{E} = \{\mathbf{e}\}$, where $\mathbf{e} = \left\langle \varepsilon_1(t), \varepsilon_2(t), \ldots, \varepsilon_l(t), \ldots, \varepsilon_{n_1}(t) \right\rangle$, $l = \overline{1, n_1}$, and the elements ε of the cortege **e** are the binary values. *The output alphabet* ∇ of the automaton A_e^c is best represented as a set of γ-vectors of the activities $\mathbf{\Gamma}(t) = \left\langle \gamma_1(t), \gamma_2(t), \ldots, \gamma_j(t), \ldots, \gamma_N(t) \right\rangle$, where $\gamma_j = \gamma_j(t)$ denoting the activity of the gene block G_j are the elements of governing gene network $S_e^c(G)$. 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 governing gene network $S_e^g(G)$, i.e., for the output channels of the automaton A_e^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 governing gene network judging, for example, by the presence (or absence) of primary transcripts. Thus, the output alphabet $\nabla = \{\mathbf{\Gamma}\}$ of the cellular automaton A_e^c represents a set of all possible words $\mathbf{\Gamma}$ of length *N*:

$\nabla = \{\mathbf{\Gamma}_1, \mathbf{\Gamma}_2, \ldots, \mathbf{\Gamma}_j, \ldots, \mathbf{\Gamma}_{2^N}\}$, where $\mathbf{\Gamma} = \mathbf{\Gamma}(t) = \left\langle \gamma_1(t), \gamma_2(t), \ldots, \gamma_j(t), \ldots, \gamma_N(t) \right\rangle$ is the γ-vector of the gene activities in the control gene network. *A set of states* Ω in the memory Ξ of the automaton A_e^c: $\Omega = \{\omega_1, \ldots, \omega_m, \ldots, \omega_M\}$, $m = \overline{(1, M)}$.

GENERAL PROPERTIES OF CELLULAR AUTOMATA

Since only living systems are treated as prototypes, one may exclude the previously introduced subnetwork ${}^{d}S_e^g(G)$ governing the processes of apoptosis (Tchuraev, 2006) from the network $S_e^g(G)$. The following ***statement about strong connection*** (1) has been derived for this case: *Any cellular automaton* $A_e^c \in A_x$ *(involved in cell ensemble* A_x*) is strongly connected*.

This statement allows *consequence* (2): *There are successions of input signals that take the cellular automaton* A_e^c *from any given state* ω_i *to any other given state* ω_j.

According to Harary (1969), from statement (1) it follows that the transition graph of the automaton A_e^c relates to Euler oriented graphs and permits the application of the theorem BEST (de Bruijn, van Aardene-Ehrenfest, Smith and Tutte), in which a formula is given for the number of Euler contours in Euler graphs. Thus, we get ***statement (3)****: In any transition graph of the cellular automaton* A_e^c *number* n_e *of Euler contours is expressed by the formula:*

$$n_e = c \prod_{j=1}^{M} (d_j-1)!,$$

where $d_j = id(\omega_j)$ is the number of vertices (states) adjacent to ω_i (semidegree of entry), and c is the common value of all cofactors of the matrix $\mathbf{M}_{cd}$ obtained by substituting $od(\omega_j)$ for main diagonal. Here $od(\omega_j)$ is the number of states (vertices) adjacent from ω_j, M is the number of states of the automaton A_e^c. As seen from this formula, number n_e of Euler contours in Moor's diagrams is strongly dependent on the number of states of the cellular automaton A_e^c and value d_i (semidegree of entry), i.e., the number of vertices (states) adjacent to ω_j. According to statement (2), the semidegree of entry $d_j = id(\omega_j) > 2$. Thus, the number of Euler contours in the transition graph of an arbitrary cellular automaton A_e^c is likely to be sufficiently high.

What is the way of interpreting this result? Let two automata A_e^c/ω_α and A_e^c/ω_β ($\omega_\alpha \neq \omega_\beta$) enter one and the same cellular ensemble A_x. If the output symbol (signal) of one automaton, A_e^c/ω_α, is the input signal for another automaton A_e^c/ω_β, i.e., $\Gamma_\alpha = \varepsilon_\beta$, *in consequence of Euler contours occurred in Moor's diagrams the automaton* A_e^c/ω_β *can transit to one of the previous states*. It should be noted that in this case the copies of automata A_e^c/ω_α transit to another state ω_j and the development of the cell ensemble will continue with conservation of the automata subpopulation of previous (non-specialized) states – "stem cells of different tissues" and "ontogenetic variability" reserve.

From statement (1) about strong connection it also follows that *any cellular automaton* A_e^c *is reversible*, i.e., always capable of being set in its initial state. Moreover, it can always be set in any predetermined state.

CELLULAR AUTOMATA IDENTIFICATION

The *identification* problem is that when an automaton is seen as a "black box" and one needs to find transition and output functions by means of measurements at its output channels. An automaton is *identifiable* if it can be identified irrespective of its initial state. It is obvious that if there is no sufficient information on the automaton, the general problem of its identification cannot be solved. By and large four laws of functioning and principles of organization have been stated for an arbitrary cellular automaton A_e^c (Tchuraev, 2006). That is why, strictly speaking, the cellular automaton A_e^c cannot be thought of as a "black box", where only output responses of the automaton to input effects are observable by an experimenter. Hence, it may be believed that an arbitrary cellular automaton A_e^c is *identifiable*.

Let us next invoke general results given in (Moor, 1956) and (Gill, 1961; 1964). In keeping with the cited studies, the process of applying input sequences to automata, observation of resultant output sequences and deduction of conclusions based on these observations are said to be an *experiment*.

According to statement 1, any cellular automaton A_e^c is strongly connected. Since for automaton A_e^c we have a predetermined number of states M, input symbols n_1 and output symbols 2^N, in compliance with (Gill, 1964) we then get the following ***statement* (4)**: *Isolated cellular automaton* A_e^c *may always be identified by a simple unconditional experiment of length* r, where

$$r \le \frac{(2M-1)(M\,2^N)^{n_1 M}}{(M-1)!} \exp\left[-\frac{M(M-1)}{2(M\,2^N)^{n_1}}\right]$$

From statement (1), according to the theorem in (Gill, 1964) **statement (5)** also follows: *Each cellular automaton* $A_e^c \in A_x$ *may be set in any predetermined state by a simple conditional experiment of length l and order d*, where

$l \le ½(M+2)(M-1)$
$d \le M$,

where M is the number of states of cellular automaton inner memory.

In order to set an arbitrarily chosen automaton ${}^{*}A_e^c/{}^{*}\omega$ to a predetermined state, particularly to ${}^{z}\omega$ ("zygotic state"), the given automaton should be isolated from adjacent automata ("to have it cultured *in vitro*").

INTERPRETATION OF CELLULAR AUTOMATON AS A FUNCTIONAL CONSTRUCTION

States $\omega_j \in \Omega$ of any cellular automaton A_e^c encode *functional* information processed in the cell governing gene/epigene network, the state ${}^{z}\omega$ therewith encodes the inherited functional information. Functional information is encoded in a cell by the pattern (CM-pattern) of different regulatory molecules and their complexes varying both in "age" and quantity as well as by spatial coordinates, when it has coding sense. As, according to the

description of the automaton A_e^c, the γ-vector value of activities $\Gamma(t)$ in neutral environment (envastat) is uniquely determined by its state, it may be stated that *the γ-vector value of activities* $\Gamma(t) = \langle \gamma_1(t), \gamma_2(t),\ldots, \gamma_j(t), \ldots, \gamma_h(t),\rangle$ *in genetic blocks of the network* $S_e^g(G)$ *under envastat conditions is uniquely determined by the CM-pattern*. From this statement and postulate C **statement (5)** follows: *Every cytotype (molecular fenotype) in any cell of a given organism x under envastat conditions is uniquely determined by the CM-pattern at previous moment of time.* There are two consequences emerging from this state.

Consequence (6). *A zygote cytotype of a given organism x under envastat conditions is uniquely determined by the CM-pattern formed in the previous generation*.

Consequence (7). *Any somatic cell cytotype of a given organism x is determined by the CM-pattern formed during ontogenesis of this generation.*

CONCLUSION

Thus, while complete identification algorithms of cellular automaton A_e^c are yet to be found, i.e., detection of its characteristic functions Φ (transitions) and Ψ(outputs), it is still possible to reveal their properties, particularly by means of graph theory.

REFERENCES

Gill A. (1961) State-identification experiments in finite automata. *Information and Control,* **4,** 132–154.

Gill A. (1964) *Introduction to the theory of finite-state machines.* Mc Graw-Hill book comp. INC. N.Y., San-Francisco, Toronto, London.

Harary F. (1969) Graph Theory. Addison-Wesley Publ. Company, London. 290 p.

Moor E.P. (1956) *Gedanken-experiments on sequential machines.* Automata Studies, Princeton University Press, Princeton, N.J.

Ratushny A.V., Likhoshvai V.A., Ignatieva E.V., Matushkin Y.G., Gorynin I.I., Kolchanov N.A. (2003) *Dokl. Russ. Acad. Nauk*., **389(**2**)**, 90–93. (In Russ.).

Tchuraev R.N., Galimzynov A.V. (2001) Modeling of actual eukaryotic control gene subnetworks with the method of generalized threshold models as the base. *Mol. Biol*., **35(**6**),** 933–939.

Tchuraev R.N. (2006) General principles of organization and laws of functioning in governing gene networks. In Kolchanov N., Hofestaedt R. (eds), *Bioinformatics of Genome Regulation and Structure II.* Springer Science + Business Media, Inc., pp. 367–377.

PART 5. COMPARATIVE AND EVOLUTIONARY GENOMICS AND PROTEOMICS

TIME SCALE OF POXVIRUS EVOLUTION

Babkin I.V.**, *Shchelkunov S.N.
State Research Center of Virology and Biotechnology "Vector", Koltsovo, Novosibirsk oblast, 633159, Russia
* Corresponding author: e-mail: babkin@vector.nsc.ru

Key words: DNA virus, Poxviridae, evolution, molecular clock, smallpox history

SUMMARY

Unlike in vertebrates and RNA viruses, the molecular clock has not been estimated so far for DNA viruses. The extended conserved central region (102 kb) of the orthopoxvirus genome and the DNA polymerase gene (3 kb) were analyzed in viruses representing several genera of the family Poxviridae. Analysis was based on the known dating of the variola virus (VARV) transfer from Western Africa to South America and previous data on phylogenetic relatedness of modern West African and South American isolates of VARV. The mutation accumulation rate was for the first time estimated for these DNA viruses at $0.9–1.2 \times 10^{-6}$ substitutions per site per year. It was assumed that poxviruses diverged from an ancestor about 500,000 years ago to form the modern species and that the ancestor of the genus *Orthopoxvirus* emerged about 300,000 years ago and give origin to the modern species about 14,000 years ago.

INTRODUCTION

The time scale of animal evolution is based on paleontological data. Comparing the nucleotide sequence for genes conserved among different animal taxa, it is possible to estimate the rate of substitutions having arisen in the given locus during the divergence of the taxa from their ancestor. The mutation rate is assumed to be constant. The resulting estimate is known as molecular clock (Russell, 1998).

In the case of viruses with single-stranded genomic RNA, the mutation rate is extremely high and the rate of virus evolution can be estimated from the accumulation of changes in nucleotide sequences of strains isolated within several years (Jenkins *et al*., 2002).

The mutation rate of DNA viruses is far more difficult to estimate. It is hardly possible to determine the molecular clock of DNA viruses by comparing the genome sequence for available isolates of one species, because mutations accumulate at a low rate and the isolation dates differ by no more than several decades. There is still no reliable dating of the divergence of virus species or genera from an ancestor of the corresponding family.

Analysis of the evolutionary relationships of DNA viruses belonging to various taxa yields phylogenetic trees lacking a time scale. Phylogenetic trees of some viruses are graphically similar to tress constructed for particular genes of the host animals (McGeoch, Cook, 1994). A hypothesis has been advanced on this evidence that viruses emerged during early evolution of host organisms and diverged into different taxa in parallel with divergent evolution of the hosts. Proceeding from this hypothesis, three modern Herpesviridae subfamilies of mammalian viruses have been assumed to originate 180–220 Myr ago (McGeoch *et al*., 1995). Yet analysis of a larger sample has shown that the phylogenetic relationships among viruses of the family do not always formally coincide with the relationships among their hosts (McGeoch, Gatherer, 2005).

We think it improper to take the same molecular clock for viruses and their natural hosts. It is more correct to assume that an ancestor of a particular virus family had initially a broad range of hosts from different taxa and that progressive specialization of viruses to different hosts took place during their long-term coevolution (Herniou *et al*., 2004). The results of such evolutionary specialization are detectable by comparing the nucleotide sequence for viral genes. Yet the molecular clock of DNA viruses is usually impossible to establish.

A unique situation is with the variola virus (VARV), which belongs to the genus *Orthopoxvirus* of the family Poxviridae. VARV (or its ancestor) caused human epidemics in Asia and Africa from long ago and in Europe from the 4th century. However, VARV was absent from America until the 16th century, when it was brought with slaves from Western Africa to South America and caused devastating epidemics among indigenous populations. More recently, variola epidemics with a low mortality were observed in South America (Shchelkunov *et al*., 2005).

The mortality during variola epidemics varied with time and geographic region in a broad range from 0.2 to 40 %. Two subtypes of VARV are distinguished accordingly: VARV major, causing epidemics with a high (from 5 to 30–40 %) mortality of infected people, and VARV minor, causing a lower mortality. South American VARV is known as VARV minor alastrim. Laboratory tests have shown that viruses of this subtype differ in several properties from African isolates of VARV minor (Shchelkunov *et al*., 2005). Moreover, VARV minor alastrim clearly differs from other VARV strains from various regions of Asia and Africa on evidence of phylogenetic analysis of individual genes (Mikheev *et al*., 2004) and extended genome regions (Shchelkunov *et al*., 2001; Babkina *et al*., 2004a). Yet the results obtained so far do not allow a timing of evolutionary changes in the VARV genome.

We were the first to introduce the time factor in phylogenetic analysis of 63 VARV genomes, which were tested for restriction fragment length polymorphism (Babkina *et al*., 2004a). South American strains proved to be closely related for West African strains. Assuming that divergent evolution of these geographically isolated variants of VARV started in the 16th century (about 400 years ago), we deduced that the West African VARV diverged from its ancestor about 1100–1300 years ago.

In this work, we applied the resulting time scale to the estimation of the poxvirus molecular clock via analyzing the conserved central region (about 102 kb) of the orthopoxvirus genome and the DNA polymerase gene (2967–3039 bp) in different genera of the family Poxviridae.

METHODS AND ALGORITHMS

Poxvirus nucleotide sequences used in this work were extracted from GenBank used in this work were extracted from GenBank and are available from authors upon request.

Phylogenetic analysis was performed and the mutation accumulation rate was estimated using the programs ClustalX v. 1.81 (Thompson *et al*., 1997), BioEdit v. 7.0.0 (Hall, 1999), SEQBOOT, CONSENSE, and DnaMLK with an optional equalizing of the probabilities of sequence input orders from the PHYLIP package v. 3.61 (Felsenstein, 1989).

RESULTS AND DISCUSSION

The family Poxviridae includes subfamilies of vertebrate (Chordopoxvirinae) and insect (Entomopoxvirinae) viruses. Chordopoxvirinae are divided into eight genera: one (*Avipoxvirus*) combining avian viruses and seven, mammalian viruses. Of the latter, viruses of the genus *Orthopoxvirus* have been studied most extensively, because this

genus includes the human pathogens VARV, monkeypox virus, and cowpox virus along with the vaccinia virus, which serves as a live vaccine against variola and other orthopoxvirus infections (Shchelkunov *et al*., 2005).

Phylogenetic analysis of data on restriction fragment length polymorphism in a large sample of VARV strains has revealed a close relationship between West African and South American isolates (Babkina *et al*., 2004a, b). Since VARV was brought from Western Africa to South America in the 16th century, we deduced that the West African variant diverged about 1100-1300 years ago from a common ancestor of modern VARV (Babkina *et al*., 2004b). With the resulting time scale of VARV evolution, we phylogenetically analyzed the extended central conserved (genus-specific) region in the sequenced genomes of several species of the genus *Orthopoxvirus*. Using the ClustalX and BioEdit programs, a 102,374-bp sequence (102 genes) was aligned for ten strains of different orthopoxvirus genes. The alignment was analyzed by the maximum likelihood method with a molecular clock, using the PHYLIP package. The significance of the resulting tree was tested by bootstrap analysis. The results showed with a high significance that the modern orthopoxvirus species separated from their ancestor about 14,000 years ago (Fig. 1).

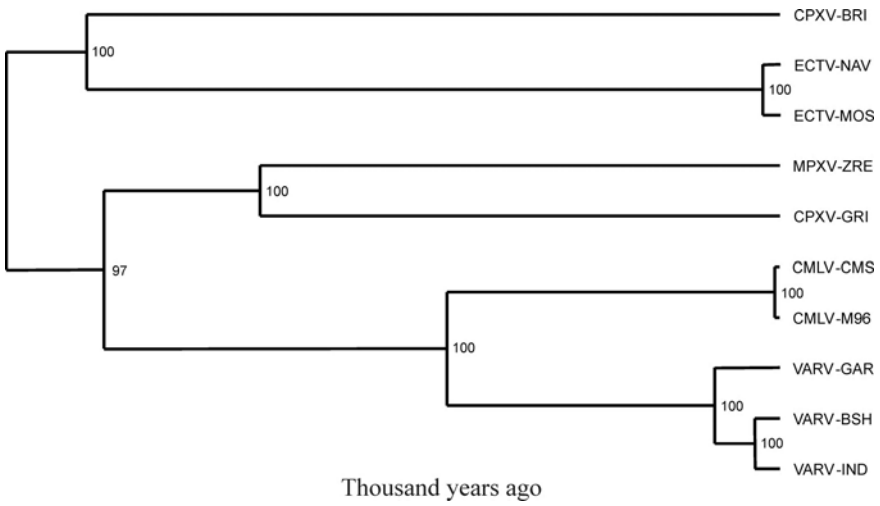


Figure 1. Phylogenetic tree constructed by analyzing the genome region highly conserved in the genus *Orthopoxvirus* and bounded by C8L and A24R according to the nomenclature of VARV strain India-1967. Here and in Fig. 2: Analysis was performed by the maximum likelihood method with a molecular clock. The significance of internal nodes is indicated.

The orthopoxvirus strains grouped by species on the phylogenetic tree (Fig. 1). The only exception was the cowpox virus: one strain clustered together with the monkeypox virus and the other, with the ectromelia virus. These findings and our previous phylogenetic analysis of the chemokine-binding protein in 85 orthopoxvirus strains (Mikheev *et al*., 2004) make it possible to assume that the cowpox virus species consists of at least two subspecies. The camelpox virus is the closest relative of VARV. The two species, each strongly specific for its host, diverged from a common ancestor (which was probably zoonotic) about 6000 years ago. The divergence coincided with the appearance of large human settlements, domestication of ungulates, and their accumulation in considerable herds, which expedited the spreading of infection among humans and animals.

It is unfeasible to compare extended genome regions for poxviruses of different genera because of considerable intergeneric differences in their organization. To date the evolutionary divergence of vertebrate poxvirus genera, phylogenetic analysis was performed with the nucleotide sequence (2967–3039) of the conserved DNA polymerase gene, using the same method as with the genome region conserved among

orthopoxviruses (Fig. 2). We analyzed nucleotide sequences of 26 strains of different vertebrate orthopoxviruses representing eight genera. The strains proved to cluster by genera. Most branches of the resulting tree were highly significant, with the exception of viruses of the genus *Yatapoxvirus*. The separation of species of the genus *Orthopoxvirus* was less reliable with the single DNA polymerase gene (Fig. 2) than with the genome region including 102 genes (Fig. 1).

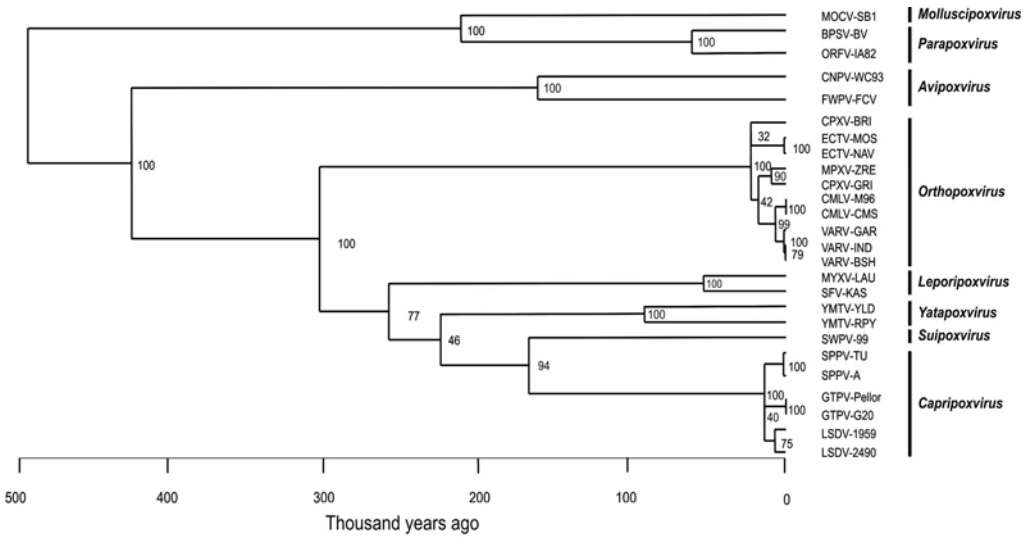


Figure 2. Phylogenetic tree constructed by analyzing the nucleotide sequence of the DNA polymerase gene in vertebrate poxviruses.

The results shown in Fig. 2 demonstrate that two evolutionary branches diverged from an ancestral virus about 500,000 years ago. One of these branches is now represented by two genera, *Parapoxvirus* and *Molluscipoxvirus*, whose DNA is characterized by a high (64.0–64.5 %) GC content. The second branch combines the other genera and is characterized by a low (25.0–43.6 %) GC content in viral DNA. Avian viruses of the genus *Avipoxvirus* diverged from mammalian viruses with a low GC content about 420,000 years ago. Although the separation of the modern orthopoxvirus species from a common ancestor was relatively recent (Fig. 1), the genus *Orthopoxvirus* diverged from the other poxivurses about 300,000 years ago (Fig. 2).

Estimation of the molecular clock showed that mutations accumulate in poxviruses at a rate of $0.9–1.2 \times 10^{-6}$ substitutions per site (nucleotide) per year. This estimate is three orders of magnitude lower than the mutation rate of viruses having single-stranded genomic RNA (Jenkins *et al*., 2002) and three to four orders of magnitude higher than the evolutionary rate of animal chromosomal genes (Russell, 1998). The poxvirus molecular clock objectively reflects the fact that alternation of generations proceeds at a far greater rate in viruses than in their hosts.

Thus, we were the first to estimate the molecular clock for DNA viruses exemplified by the family Poxviridae. Our data call into question the hypothesis that modern viruses originate from ancestors that emerged and started divergent evolution hundreds or tens of million years ago, simultaneously with the development of new animal taxa (McGeoch *et al*., 1995).

REFERENCES

Babkina I.N., Babkin I.V., Marennikova S.S., Sandakhchiev L.S., Shchelkunov S.N. (2004a) Comparative restriction enzyme analysis of the genome in variola virus strains from the Russian collection. *Mol. Biol.*, **38**, 429–436.

Babkina I.N., Babkin I.V., Le U, Ropp S., Kline R., Damon I., Esposito J., Sandakhchiev L.S., Shchelkunov S.N. (2004b) Phylogenetic comparison of the genomes of different strains of variola virus. *Dokl. Biochem. Biophys.*, **398**, 818–822.

Jenkins G.M., Rambaut A., Pybus O.G., Holmes E.C. (2002) Rates of molecular evolution in RNA viruses: a quantitative phylogenetic analysis. *J. Mol. Evol.*, **54,** 152–161.

Herniou E.A., Olszewski J.A., O'Reilly D.R., Cory J.S. (2004) Ancient coevolution of baculoviruses and their insect hosts. *J. Virol.*, **78,** 3244–3251.

McGeoch D.J., Cook S. (1994) Molecular phylogeny of the alphaherpesvirinae subfamily and a proposed evolutionary timescale. *J. Mol. Biol.*, **238,** 9–22.

McGeoch D.J., Cook S., Dolan A., Jamieson F.E., Telford E.A.R. (1995) Molecular phylogeny and evolutionary timescale for the family of mammalian herpesviruses. *J. Mol. Biol.*, **247,** 443–458.

McGeoch D.J., Gatherer D. (2005) Integrating reptilian herpesviruses into the family Herpesviridae. *J. Virol.*, **79,** 725–731.

Mikheev M.V., Feshchenko M.V., Shchelkunov S.N. (2004) Phylogenetic analysis of chemokine-binding protein gene from orthopoxviruses. *Mol. Gen. Mikrobiol. Virusol.*, **1**, 29–36.

Russell P. (1998) *Genetics*. 5th ed. Melno Park, California: Addison Wesley Longman Inc.

Shchelkunov S.N., Marennikova S.S., Moyer R.W. (2005) *Orthopoxviruses Pathogenic for Humans*. Berlin, Heidelberg, New York: Springer.

Shchelkunov S.N., Totmenin A.V., Babkin I.V., Safronov P.F., Ryazankina O.I., Petrov N.A., Gutorov V.V., Uvarova E.A., Mikheev M.V., Sisler J.R., Esposito J.J., Jahrling P.B., Moss B., Sandakhchiev L.S. (2001) Human monkeypox and smallpox viruses: genomic comparison. *FEBS Lett.*, **509**, 66–70.

Thompson J.D., Gibson T.J., Plewniak F., Jeanmougin F., Higgins D.G. (1997) The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucl. Acids Res.*, **24**, 4876–4882.

BURSTS OF NON-SYNONYMOUS SUBSTITUTIONS IN HIV-1 PHYLOGENETIC TREE REVEAL INSTANCES OF POSITIVE SELECTION AT CONSERVATIVE PROTEIN SITES

***Bazykin G.A.*[*1], *Dushoff J.*[1], *Levin S.*[1], *Kondrashov A.*[2]**
[1] Dept. of Ecology and Evolutionary Biology, Princeton University, Princeton, NJ 08650, USA;
[2] National Center for Biotechnology Information, NIH, Bethesda, MD 20894, USA
[*] Corresponding author: e-mail: gbazykin@princeton.edu

Key words: evolution, positive selection, HIV-1, bursts of substitutions

SUMMARY

Motivation: Darwinian positive selection can cause fixation of novel alleles in population. The conventional methods of identification of positive selection rely on excess of non-synonymous over synonymous substitutions. This approach is adequate for identification of selection affecting lineage for a substantial period of time. However, it is biologically realistic to expect that some selection events can be transient, affecting only one to a few successive substitution events at an otherwise conservative site.

Results: We analyzed pairs of successive substitutions within amino acid position in evolution of HIV-1 genes. Successive non-synonymous substitutions tend to be much more clumped than the random expectations and than synonymous substitutions. Strikingly, this effect is strongest in sites with low overall rate of non-synonymous substitutions. The observed bursts of amino acid-changing substitutions can not be explained by mutation biases, and are, therefore, due to positive selection.

INTRODUCTION

Positive selection which favors new alleles drives adaptive evolution and, thus, is of paramount importance (Fisher, 1930). However, positive selection always occurs on the background of pervasive negative selection which maintains *status quo* (Williams, 1966). Even at the simplest level of DNA and protein sequences, disentangling the two remains a major challenge.

We pursue a novel approach to detecting positive selection. Due to the nature of the genetic code, replacing amino acid X with amino acid Z often requires two or even three non-synonymous substitutions, since in only 75 out of 190 unordered amino acid pairs the members can be converted into each other by a single nucleotide substitution. Thus, when positive selection favors a particular amino acid replacement, in a large fraction of cases a clump of two or three successive nucleotide substitutions should occur (Gillespie, 1984), even if most of the time the site evolves slowly, due to negative selection.

Comparison of rat, mouse, and human orthologous proteins demonstrated that, at codon when rat and mouse differ by two nonsynonymous substitutions, both substitutions tend to occur, after rat-mouse divergence, in the same lineage, either rat or mouse (Bazykin *et al*., 2004). This result indicates that clumps of successive nonsynonymous substitutions may be common. However, since rat-and mouse are tightly related to each

other, codon at which they differ by two nucleotide substitutions are rare and must mostly come from the subset of rapidly-evolving codons.

Here we analyze the evolution of 4 genes of HIV-1, which collectively encode 10 proteins, using sets of many hundreds of genomes. These data make it possible to look for clumps of nonsynonymous substitutions both at rapidly evolving and at slowly evolving amino acid sites.

MATERIALS AND METHODS

Sequences and phylogenies. Alignments of nucleotide sequences of all full-length *env, gag, pol,* and *nef* protein coding regions from HIV-1 genomes of subtypes A-H were taken from the 2003 Los Alamos National Laboratory HIV-1 sequence database (Korber *et al*., 2000a). For each gene, a maximum parsimony tree was constructed by PAUP. Resulting trees were rooted using the consensus of consensus sequences for each subtype (Korber *et al*., 2000b). All the analysis was conducted using a set of Bioperl-based scripts (Stajich *et al*., 2002) which are available from the authors upon request.

Analysis of nucleotide substitutions. We used maximum parsimony to reconstruct the states of the codons at all internal nodes within each phylogeny. Then, for each codon, we inferred the edges of the tree at which each single-nucleotide substitution occurred, as follows. If a pair of successive nodes within a phylogeny differed at a nucleotide position, we assumed that exactly one nucleotide substitution occurred on the edge connecting these nodes. If the codons at successive nodes differed at more than one nucleotide site, the numbers of synonymous and non-synonymous substitutions were averaged over all possible orders of substitution events and rounded to the integer. For each codon, we estimated the number of non-synonymous substitutions and the number of synonymous substitutions on the whole tree.

For each codon, we treated synonymous and non-synonymous substitutions separately, so that synonymous substitutions were ignored when non-synonymous substitutions were considered and vice versa. If successive substitutions occurred along the path from the tree root to a leaf, for each substitution A, except the first one, the preceding substitution A' at the same codon could be uniquely determined. We assume that two substitutions at a codon form a pair if there is a path from the root of the tree to at least one of the leaves, such that both the substitutions belong to this path, and there are no other substitutions on this path between them. In particular, two substitutions that occurred at the same codon on the same edge, revealed by the codons at two successive nodes differing from each other at two nucleotide sites, always constitute a pair. Substitutions A' and A can occur either at the same site or at different sites of the codon.

We estimate the distance l between A' and A as the sum of the lengths of edges between them, assuming that substitutions always occur at the middle of an edge. If A' and A occurred within the same edge, $l = 0$ for them. For each codon, we calculated the distances within the pairs of successive synonymous and of non-synonymous substitutions, and compared them with the distances obtained in simulated evolution for the same total number of substitutions on the phylogeny.

Simulations of sequence evolution. To find the expected numbers of pairs of successive substitutions at a codon and the distance between substitutions in such pairs, we simulated random substitution events in each codon on the phylogeny. Simulations were then done for each amino acid position in each gene separately. First, we counted the total number of substitutions of particular type (synonymous or non-synonymous) at the given codon on the phylogenetic tree. Next, in each of the 10,000 simulation trials, we distributed the same number of events randomly over the edges of the actual phylogenetic tree, with the probabilities weighted by the lengths of the corresponding branches. We then counted the numbers of pairs and distances within pairs of successive substitutions in simulation.

RESULTS

We built the phylogenetic trees of 343, 218, 193, and 674 full-length sequences of *env*, *gag*, *pol*, and *nef* genes from HIV-1 genomes. Let us consider how non-synonymous substitutions are distributed relatively to each other. Among the 7424 observed pairs of successive non-synonymous substitutions, 3022 (40.7 %) were reversals, 836 (11.3 %) were non-reversing substitutions at the same site, and 2108 (28.4 %) substitutions occurred in different nucleotides of the same codon.

Non-synonymous substitutions display reduced evolutionary distances between the members of a pair, relative to random expectation and to the distances between successive synonymous substitutions. In contrast, the average distance between successive synonymous substitutions was close to that predicted in the simulation (Fig. 1). Clumping of successive non-synonymous substitutions is the strongest at codons with the smallest total number of non-synonymous substitutions, i.e., the ones that are highly conserved (Fig. 1).

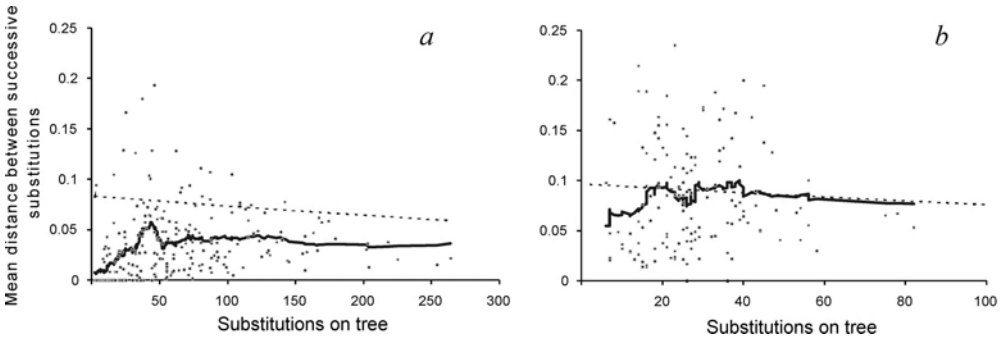


Figure 1. Mean distance between successive non-reversing substitutions, plotted against the total number of non-synonymous (*a*) and synonymous (*b*) substitutions in each site of the *env* HIV-1 gene. Solid lines indicate mean distances between successive substitutions in each sliding 30 site window. Dashed lines indicate mean distances in simulations.

DISCUSSION

Our results indicate that while the distribution of synonymous substitutions over the phylogenetic tree is consistent with the molecular clock model, the non-synonymous substitutions occur in bursts, forming pairs of rapid successive substitutions. This clumping of non-synonymous substitutions is not an artifact of phylogenetic reconstruction, of mutations or sequencing errors spanning multiple adjacent nucleotides in some of the sequences, or of our approach to inference of positions of individual substitutions (analysis not shown).

Therefore, the non-synonymous substitutions actually tend to occur in bursts. Although different codons have intrinsic differences in the rate of non-synonymous evolution due to the structure of the genetic code and peculiarities of the substitution matrix, these differences can not explain the observed clumping (data not shown). Apparently the only explanation of the clumping of amino acid-changing nucleotide substitutions that we are left with is episodes of positive selection alternating with periods of negative selection acting on individual codons.

Bursts of substitutions can be caused by a single environmental change that creates a new fitness landscape for the given locus (Gillespie, 1984). Even at a single codon, up to three successive substitutions can be facilitated by selection after a single change in fitness due to the structure of the genetic code. Indeed, after the change of fitness landscape, the new preferred amino acid can be reachable by no less than two or even

three nucleotide substitutions, and each of these substitutions can be facilitated by selection. The observed clumping of successive substitutions in a single nucleotide can be explained by assuming that the intermediate codon has lower fitness than the final variant.

We observe strongest clumping of non-synonymous substitutions in the most conservative codons, with the lowest total number of substitutions. Apparently, positive selection plays a major (and previously unobserved) role in the evolution of important and generally conservative amino acids which cannot be replaced by random drift. New substitutions that do occur in this site can also be expected to be functionally important, and get fixed rapidly under positive selection. Our results indicate a new set of sites in HIV-1 proteins, most of which are conservative, that evolve under positive selection.

ACKNOWLEDGEMENTS

GB gratefully acknowledges fellowships from the Pew Charitable Trusts award 2000-002558 and the Burroughs Wellcome Fund award 1001782, both to Princeton University. We thank Mikhail Gelfand, Sergey Kryazhimskiy and Joshua Plotkin for valuable discussions.

REFERENCES

Bazykin G.A. *et al*. (2004) Positive selection at sites of multiple amino acid replacements since rat-mouse divergence. *Nature,* **429,** 558–562.

Fisher R.A. (1930) *The Genetical Theory of Natural Selection.* The Clarendon Press, Oxford.

Gillespie J. (1984) Molecular evolution over the mutational landscape. *Evolution,* **38,** 1116–1129.

Korber B. *et al*. eds. (2000a) HIV Immunology and Sequence Databases, Los Alamos National Laboratory, Los Alamos, NM. Available at http://lanl.hiv.org

Korber B. *et al*. (2000b) Timing the ancestor of the HIV-1 pandemic strains. *Science,* **288,** 1789–1796.

Stajich J.E. *et al*. (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.,* **12,** 1611–1618.

Williams G.C. (1966) *Adaptation and natural selection; a critique of some current evolutionary thought.* Princeton University Press, Princeton.

MICRO- AND MINISATELLITES IN HUMAN GENOME, TandemSWAN SOFTWARE IN USE

***Boeva V.A.*[*1], *Makeev V.J.*[2]**
[1] Moscow State University, Department of bioengineering and bioinformatics, Moscow, Russia;
[2] GosNIIGenetika State Research Center, Moscow, Russia, e-mail: makeev@genetika.ru
[*] Corresponding author: e-mail: valeyo@imb.ac.ru

Key words: tandem repeats, microsatellites, minisatellites, human genome

SUMMARY

Motivation: Eukaryotic genomes contain many various types of regular structure. One of them is tandem repeats. They are classified into microsatellites (or simple repeats) and minisatellites. Their reproduction mechanisms are different. The hypothesis we tried to prove is that some part of minisatellites is originated from microsatellites during the evolution.

Results: Using our program TandemSWAN we identified tandem repeats in human genome (17th UCSC release). Comparison of our repeat annotation with that obtained by TRF/RepeatMasker showed that our database contains at least 20 % of repeats not annotated before. The whole database can be found at our website. We studied the distribution of micro- and minisatellites found by TandemSWAN and observed the presence of two classes of minisatellites: those originated from microsatellites (their period length seldom exceeds 25 bp) and real long minisatellites, for which it is difficult to identify the smaller sub-period.

Availability: http://bioinform.genetika.ru/.

INTRODUCTION

The recent study (Smit *et al*., 2004) shows that more than 50 % percent of human genome can be masked as tandem repeats (micro- and minisatellites) or as low complexity regions. In biological literature one denotes by *minisatellites* tandem repeats with the length of the repetitive motif from 6 to ~100 and by *microsatellites* repeats with smaller period length, usually from 2 to 6. They are considered to have different formation and multiplication mechanisms. Using our TandemSWAN software we scanned the human genome (the 17th UCSC release, Kent *et al*., 2002) in order to find all significant micro- and minisatellites, to assess their period length distribution, and to make some observations on their nature and possible origin.

METHODS AND RESULTS

To identify tandem repeats we used the TandemSWAN (Boeva *et al*., 2006) program which searches for tandem repeats with evaluation of their statistical significance. The parameters used are: maximal period length, 70 bp; significance level, 3; the “Mask” mode. In contrast to other popular programs searching for tandem repeats, such as RepeatMasker (Smit *et al*., 1996–2004), TRF (Benson, 1999), or mreps (Kolpakov *et al*.,

2003), TandemSWAN can simultaneously identify both short- and large period repeats with approximately the same fuzziness and allows one to evaluate the statistical significance of the repeats found. The fact that TandemSWAN uses efficiently the statistical significance of repeats found is essential for our purposes: it allows us to obtain the set of the most significant repeats in DNA sequences with the appropriate period identification in controversial cases. More detail on the comparison of TandemSWAN to other algorithms can be found in (Boeva *et al*., 2006). The total set of repeats found by TandemSWAN covers approximately the 30 % of human genome. But as not all of these repeats are significant, for the further study we selected tandem repeats with the –log P-value (the statistical significance) greater than 15.

First, we considered tandem repeats with the statistical significance value greater than 15 and with period lengths up to 23 bp. (Fig. 1). We observed the distinct peaks at periods multiples of 4. Interestingly, this disagrees with our previous study of *D. melanogaster* genome (Boeva *et al*., 2006) where we have found peaks at periods multiples of 6. Our hypothesis explaining the origin of multiple tandem repeat periods is that this could be the result of the following process: a microsatellite (period length ≤ 6) duplicates several times, after that (or during that) some of its units mutate and later on it is multiplied using with the minisatellite mechanism. So the result is a minisatellite but with a vestigial smaller subperiod. To test this hypothesis we did some postprocessing of repeats found in order to identify smaller repeated motifs, possibly more degenerate. As a result (Fig. 2 and Table 1) we have seen that minisatellites with small lengths of a repeated unit (up to 23) often originate from microsatellites. For human genome the effect is brighter for the repeats with the length of the repeated motif equal to 4 bp.

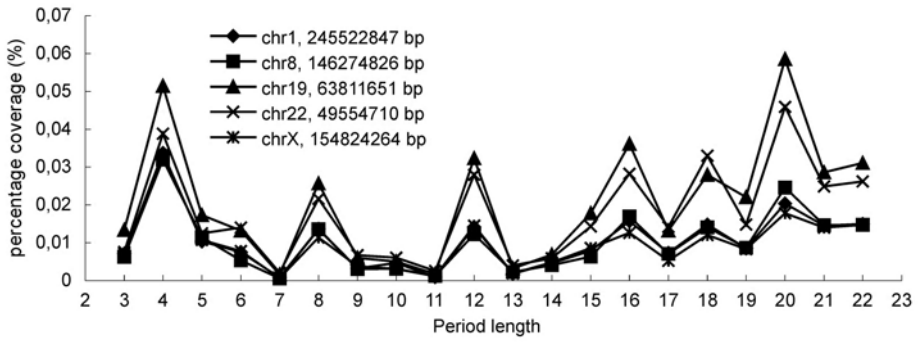


Figure 1. Total percentage coverage by tandem repeats with various period lengths of various chromosomes of human genome. Filtration by statistical significance value >15.

Apart from microsatellites and microsatellite-derived minisatellites, whose substantial part was identified by TRF/RepeatMasker, in 17th UCSC human genome we have found a great number of fuzzy minisatellites with period basically greater than 25. Only 57 % of repeats found by TandemSWAN have been masked by TRF/Repeat Masker. The total lengths of minisatellites found by TandemSWAN in the Chr22 with period length from 25 to 70 and their part masked by TRF/Repeat Masker in the annotation of human genome (the 17th UCSC release) are shown in Fig. 3. Examples of vestigial minisatellites which were not masked by TRF/Repeat Masker in the annotation of human genome (the 17th UCSC release) are shown in Table 2. The whole database can be found at /www.bionform.genetika.ru/.

Table 1. Examples of minisatellites originated from microsatellites (length of repeated unit up to 6 b.p.)

(ccag)n on the chr22	(ggat)n on the chr1	(tcatc)n on the chr22	(gccac)n on the chr22
catccaagccagccaag	tggatggatgtt	tcatttcatctca	gccatgccac
tcagccagccagccaag	tgaatggatagg	tcagttcatctca	gccatgccac
ccagccaagccagccag	tggatagatagg	tcatttcatctca	gccacgccac
ctagccaagccacccag	tggatggatgga	tcatctcatctca	accatgccaa
cca	ttggtggaggg	tcatttcatctca	gccacgccac
(15422147-5422217)	(47404286-47404344)	tcact	accatagcac
		(20965525-20965594)	accatgccac
			(42815130-4281519)

Table 2. Examples of vestigial minisatellites which were not masked by TRF/RepeatMasker in the annotation of human genome (the 17th UCSC release)

chr 22, 43460819-43461054, period length 36	chr 22, 19862193-19863098, period length 27
CATATGGGGATGCTCCCACAGCACAGAGA GGTGCCC	CATCGCTAACGAGGCCGCCGACAAGGG
ATCATATGGAGATGCTCCCACTGCACAGAT ACTCCC	CATCGCCAACGAGGATGCCGCCCACGG
ATTGCACAGATACTCCCACAGCACAGAGA GGTGCCC	CATCGCCAACGAGTACGCCGCCCACGG
ATCATATGGGGATGCTCCCACTGCACAGAT ACTCCC	CATCGCCAGCGAGGACGCCGCCCACGG
ATTGCACAGATACTCCCACCGCACAGAGA GGCACCC	CATCGCCAGCGAGGACGCCGCCCAGGG
ATGATATGGGGATGCTCCCACTGCACAGAT GCTCCC	CATCGCCAGCGAGGACGCCGCCCAGGG
ACGGCACAGAAAGGCACCCA	CATCGCCAACGAGGACACCATCCAGGG
	CATCGCCAAGGAGTACGCCGTCCACGG...
chr 22, 17196266-17196394, period length 43	
CAGTTAAAGCTCCAGCAGAAACCAGATTGGATGTTCCAGCTCC	
CAGAGAAGACTCTATCACAGTGCACATAAGGAGCACCAACGGA	
CCTATCGATGTCTATTTGTGAGAAGTGGAGCAGGGTCAGACCA	
chr 22, 23263434-23263553, period length 40	
CCCAGCAGCCCCAAGCCTGGGCCAGGGCATCCTGAGCGGG	
CCCGGCTGGGGTGGGGATGGGGTCCGAGGGCCTAGGCGCA	
CCAGGCTCCTGCTGTCAGCTGTCCAAGTAGACAAAGAGGA	

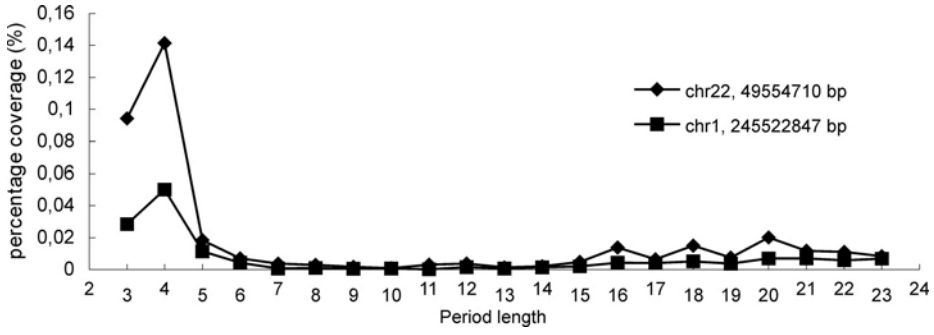


Figure 2. Results after postprocessing. Total percentage coverage by tandem repeats with various period lengths of chromosome 22. Filtration by statistical significance value >15.

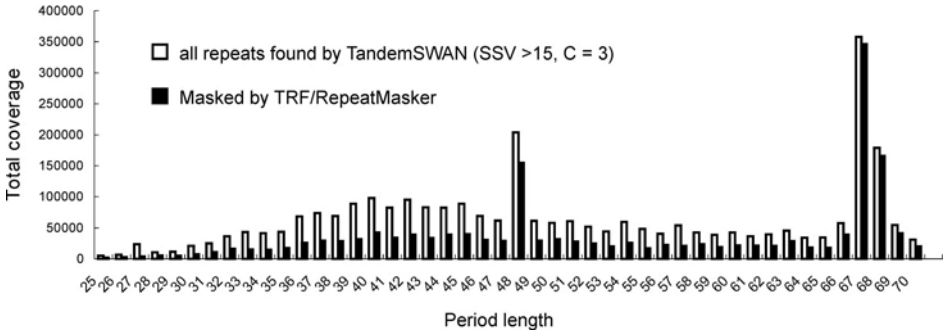


Figure 3. New repeats found on Chr22 of human genome.

ACKNOWLEDGEMENTS

This study has been supported by French Program EcoNet-08159PG, INTAS grant 04-83-3994, Russian State Contract 02.434.111008, RFBR grant 04-04-49601, Fogerthy RO3 TW005899-01A1 program, Russian Academy of Science Presidium Program in Molecular and Cellular Biology, project #10, and Ludwig Institute of Cancer Research Grant CRDF GAP RBO-1268.

REFERENCES

Benson G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucl. Acids Res*., **27**, 573–578.

Boeva V., Regnier M., Papatsenko D., Makeev V. (2006). Short fuzzy tandem repeats in genomic sequences, identification, and possible role in regulation of gene expression. *Bioinformatics*, **22**(6), 676–684. /http://bioinform.genetika.ru/projects/swan/www/

Kent W.J. *et al*. (2002). The Human Genome Browser at UCSC. *Genome Res.*, **12**(6), 996–1006. /http://genome.ucsc.edu/

Kolpakov R., Bana G., Kucherov G. (2003) mreps: efficient and flexible detection of tandem repeats in DNA. *Nucl. Acids Res.*, **31**, 3672–3678.

Smit A.F.A., Hubley R., Green P. *RepeatMasker Open-3.0*, 1996–2004. /http://www.repeatmasker.org/.

SEARCH FOR MULTI-SNP DISEASE ASSOCIATION

Brinza D., Perelygin A., Brinton M., Zelikovsky A.*
Georgia State University, Atlanta, GA, USA
* Corresponding author; e-mail: alexz@cs.gsu.edu

Key words: algorithm, disease association, SNP, genotypes

SUMMARY

Motivation: Recent improvements in the accessibility of high-throughput genotyping have brought a great deal of attention to association studies for common complex diseases. The two-loci analysis was recently developed (Marchini *et al*., 2005) but multi-loci analyses are expected to find even stronger disease associations.

Results: A novel combinatorial method for finding disease-associated multi-SNP combinations was developed. Multi-SNP combinations significantly associating with diseases were found. For Crohn's disease data (Daly *et al*., 2001), a few associated multi-SNP combinations with multiple-testing-adjusted to $p < 0.05$ were found, while no single SNP or pair of SNPs showed significant association. For a dataset for an autoimmune disorder (Ueda *et al*., 2003), a few previously unknown associated multi-SNP combinations were found. For tick-borne encephalitis virus-induced disease, a multi-SNP combination within a group of genes showing a high degree of linkage disequilibrium significantly associated with the severity of the disease was found.

Availability: http://alla.cs.gsu.edu/~software/DACS.

INTRODUCTION

Analysis of variation in suspected genes in disease and nondisease individuals is aimed at identifying SNPs with considerably higher frequencies among the disease individuals than among the nondisease individuals. Successful (as well as unsuccessful) searches for SNPs with statistically significant associations have recently been reported. Although common diseases can be caused by combinations of several unlinked gene variations, most searches are done on a SNP-by-SNP basis. In this paper, we address the computational challenge of searching for such multi-gene causal combinations.

False-discovery rates are usually very high for genome-wide searches. Although only statistically significant SNPs (with a $p < 0.05$ for frequency distribution) are reported, frequently these findings are not reproducible because the computed p-values are not adjusted for multiple testing. The standard Bonferroni adjustment is overly pessimistic; therefore, we adjusted for multiple testing by using a more accurate randomization method.

Formally, the computational problem is as follows. Given a population of disease and nondisease n genotypes or haplotypes with values of m SNPs, find all multi-SNP combinations with multiple-testing adjusted to $p < 0.05$ for the frequency distribution. We show that this problem is computationally feasible using the proposed novel searching techniques.

METHODS AND ALGORITHMS

The search for disease-associated multi-SNP combinations, i.e., with the p-value of the frequency distribution below 0.05, among all possible combinations can be done by *Exhaustive Search* (ES) (e.g., Marchini *et al*., 2005). Since ES checks all 1-SNP, 2-SNP, ..., *m*-SNP combinations, its runtime is $O(n3^m)$ making it unfeasible even for small numbers of SNPs *m*. Further we searched only among multi-SNP combinations with k < 3 SNPs. We refer to k as the *search level* of the exhaustive search. In order to reduce the runtime of the exhaustive search, we propose to decrease the size of the input data set by extracting informative SNPs (*indexing SNPs*) from which one can reconstruct all other SNPs. In our experiments, we used a multiple linear regression based tagging method (He, Zelikovsky, 2006). The tradeoff between the number of chosen indexing SNPs and quality of reconstruction requires choosing the maximum number of index SNPs that can be handled by ES in a reasonable computational time. ES on indexing SNPs will be further referred as *Indexing Exhaustive Search* (IES).

Another way to fight extensive computations is to apply a faster search. Our new search method can find disease-associated multi-SNP combinations consisting of large numbers of SNPs and small search levels *k*. The new method is based on the notion of closure. Let *C* be a multi-SNP combination and let *snp*(*C*) be the subset of SNPs with their values defining C. All individuals containing *snp*(*C*) are partitioned into two subsets: *dis*(*C*) consisting of individuals with disease and *nondis*(*C*) consisting of individuals without disease. We search for *C*'s with larger disease frequency *dis*(*C*) and lower nondisease frequency *nondis*(*C*). Sometimes, the size of *nondis*(*C*) can be decreased while keeping *dis*(*C*) unchanged using the following closure operator. The set *dis*(*C*) can have more SNP values in common than in *snp*(*C*). Closure of *C* is a multi-SNP combination *C'* with *snp*(*C'*) equal to *snp*(*C*) extended with such SNPs. Obviously, *dis*(*C'*) = *dis*(*C*) while *nondis(C'*) $\subseteq$ *nondis*(*C*). The proposed *Combinatorial Search* (CS) finds the disease-associated multi-SNP combinations among closures of all j-SNP combinations (j = 1..k, k < m). The corresponding *search level* is the number of SNPs *k* in multi-SNP combinations for which a closure is found. Because of the disease-closure, the same level of searching using a combinatorial search finds better associations than an exhaustive search. CS on indexing SNPs will be further referred as *Indexing Combinatorial Search* (ICS).

IMPLEMENTATIONS AND RESULTS

In our implementation, we avoid checking of those multi-SNP combinations which can not lead to statistically significant ones. Additionally, we never recheck any combinations after disease-closure that will form already checked combinations. The resulting implementation is several times faster than the exhaustive search.

Results from four methods used to search disease-associated multi-SNP combinations are reported for the following datasets. The first dataset was derived from human Chromosome 5q31, which may contain a genetic variant responsible for Crohn's disease, by genotyping 103 SNPs for 144 disease and 243 nondisease individuals (Daly *et al*., 2001). The second dataset consisted of 108 SNPs sequenced from 330 kb of human DNA containing the genes, CD28, CTLA4 and ICONS, that were previously shown to be related to an autoimmune disorder, from 384 disease and 652 nondisease individuals (Ueda *et al*., 2003). The tick-borne encephalitis virus-induced dataset consists of 41 SNPs genotyped from DNA of 21 patients with severe tick-borne encephalitis virus-induced disease and 54 patients with mild disease. The missing genotypes were inferred using 2SNP software (Brinza, Zelikovsky, 2006).

The four methods compared are *Exhaustive Search* (ES), *Indexed Exhaustive Search* (IES(N))-ES on the indexed datasets obtained by extracting N indexed SNPs, *Combinatorial Search* (CS), *Indexed Combinatorial Search* [ICS(N)]-CS on the indexed datasets. Each method has been applied only to the search levels of 1 and 2. All experiments were run on a Processor Pentium 4 3.2Ghz, RAM 2Gb, OS Linux. The runtime is given in the last column of Table 1.

Table 1. Methods for searching disease associated multi-SNPs combinations

Search level	Search method	MT-unadjusted p corresp. to adjusted $p = 0.05$	SNP combination with minimum p-value			# of SNP combin. with MT-adjusted $p < 0.05$	Run-time msec
			expos. freq.	unexp. freq.	unadjust. p-value		
Dataset of Crohn's disease (Daly *et al*., 2001)							
1	ES	1.6×10^{-3}	0.31	0.16	1.8×10^{-3}	0	900
	IES(30)	3.9×10^{-3}	0.30	0.16	4.7×10^{-3}	0	500
	CS	5.1×10^{-5}	0.30	0.11	2.0×10^{-5}	2	1000
	ICS(30)	2.2×10^{-3}	0.30	0.14	4.6×10^{-4}	1	600
2	ES	1.9×10^{-5}	0.30	0.13	3.1×10^{-4}	0	15000
	IES(30)	1.0×10^{-4}	0.31	0.14	4.4×10^{-4}	0	1000
	CS	1.5×10^{-6}	0.17	0.02	6.5×10^{-7}	2	7000
	ICS(30)	5.0×10^{-5}	0.17	0.04	3.7×10^{-5}	1	400
Dataset of autoimmune disorder (Ueda *et al*., 2003)							
1	ES	1.3×10^{-3}	0.43	0.28	1.1×10^{-4}	2	1000
	IES(30)	3.1×10^{-3}	0.43	0.28	1.1×10^{-4}	4	600
	CS	1.8×10^{-4}	0.43	0.28	9.2×10^{-5}	2	1100
	ICS(30)	1.6×10^{-3}	0.43	0.28	1.1×10^{-4}	4	600
2	ES	2.7×10^{-6}	0.25	0.12	1.5×10^{-6}	2	30000
	IES(30)	8.0×10^{-5}	0.25	0.12	1.5×10^{-6}	9	3000
	CS	1.1×10^{-6}	0.16	0.06	8.5×10^{-7}	3	20000
	ICS(30)	4.7×10^{-5}	0.25	0.12	1.1×10^{-6}	10	1000
Dataset of tick-borne encephalitis virus-induced disease							
1	ES	6.1×10^{-3}	0.33	0.07	1.5×10^{-2}	0	80
	IES(20)	9.4×10^{-3}	0.33	0.07	1.5×10^{-2}	0	30
	CS	4.8×10^{-4}	0.33	0	1.3×10^{-4}	1	84
	ICS(20)	8.1×10^{-4}	0.33	0.02	8.1×10^{-4}	1	35
2	ES	2.5×10^{-4}	0.29	0	4.8×10^{-4}	0	820
	IES(20)	1.3×10^{-4}	0.29	0	4.8×10^{-4}	0	100
	CS	4.3×10^{-5}	0.33	0	1.3×10^{-4}	0	600
	ICS(20)	1.3×10^{-4}	0.29	0	4.8×10^{-4}	0	76

The relative qualities of the searching methods are compared using the number of statistically significant multi-SNP combinations found (Table 1, column 7). The statistical significance was adjusted to multiple testing and the adjusted 0.05 threshold is shown (Table 1, column 3). In the 4th, 5th and 6th columns, we give the frequencies of the best multi-SNP combination among disease and nondisease populations and the unadjusted p-value, respectively.

DISCUSSION

Comparing indexed counterparts with ES and CS shows that indexing is quite successful. Indeed, the indexed searches found the same multi-SNP combinations as the non-indexed searches for the second and third data sets but were much faster and the multiple-testing adjusted 0.05-threshold was higher and easier to meet.

Comparing the CS with the ES counterparts is advantageous to the former. Indeed, for the Crohn's disease data (Daly *et al*., 2001), the ES on the first and second search levels is unsuccessful while the CS finds several statistically significant multi-SNP combinations. Similarly, for the tick-borne encephalitis virus-induced disease data, the CS and ICS(20) found a significant association on the first level while no association was found by the ES or IES(20). For the autoimmune disorder data (Ueda *et al*., 2003), the CS found many more statistically significant multi-SNP combinations then the ES.

In addition, we have developed a new disease susceptibility prediction (DSP) method based on CS. In a leave-one-out test for the tick-borne encephalitis data, the accuracy of DSP is 90 % which proves that the data contain a well-defined border between severe and mild forms. The accuracy of DSP is around 85 % for the other two datasets which is significantly higher than the accuracy of previously known methods. These results show that the combinatorial disease-association search is more powerful than the existing methods when applied to disease susceptibility prediction.

We conclude that the proposed indexing approach and the combinatorial search method are very promising techniques for searching for statistically significant diseases-associated multi-SNP combinations and disease susceptibility prediction.

ACKNOWLEDGEMENTS

DB was supported by GSU Molecular Basis of Disease Fellowship, AZ was supported by NIH Award 1 P20 GM065762-01A1 and US CRDF Award MOM2-3049-CS-03, and AP was supported by CDC grant R01 CI000216.

REFERENCES

Brinza D., Zelikovsky A. (2006) 2SNP: scalable phasing based on 2-SNP haplotypes. *Bioinformatics*, **22(**3**)**, 371–373.

Daly M., Rioux J., Schaffner S., Hudson T., Lander E. (2001) High resolution haplotype structure in the human genome. *Nature Genet.*, **29**, 229–232.

He J., Zelikovsky A. (2006) Tag SNP selection based on multivariate linear regression, *Proc. of the Intern. Conf. on Computational Science (ICCS 2006) (to appear)*.

Marchini J., Donnelley P., Cardon L.R. (2005) Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nature Genet.*, **37**, 413–417.

Ueda H., Howson J.M.M. *et al*. (2003) Association of the T cell regulatory gene CTLA4 with susceptibility to autoimmune disease. *Nature*, **423**, 506–511.

THE MODELS OF POPULATION DYNAMICS AS TOOL FOR STUDYING OF GENETIC POLYMORPHISM OF BAIKALIAN POLYCHAETS

Bukin Yu.S.**, Pudovkina T.A.
Limnological Institute, SB RAS, Irkutsk, 664033, Russia
* Corresponding author: e-mail: bukinyura@mail.ru

Key words: population dynamics, DNA, genetic polymorphism, mitochondria, population genetics, spatial structure of population

SUMMARY

In some cases it is difficult to interpret accurately the results obtained with molecular genetic analysis. One of such examples is the case of fresh-water polychaets from Lake Baikal. The observable picture of genetic differentiation in this group of organisms does not correspond to the standard Wright's (Wright, 1943) model "Isolation by Distance". To obtain the more correct explanation, the methods of population dynamics simulating the processes of neutral molecular evolution are used. With this model we estimated possible results of different scenarios of migration and interaction of organisms and determined stationary state of neutral DNA polymorphism. If DNA polymorphism in the model fits the pattern observed, it allows us to make assumption that causes of genetic polymorphism in the model and in the real population are similar.

INTRODUCTION

In this work we try to explain the mechanisms generating the pattern of genetic polymorphism in populations of organisms by means the population dynamics. This necessity has arisen since for some groups of organisms where it is difficult to propose adequate explanation for the pattern observed. One of such groups of organisms is Baikalian polychaete genus *Manayunkia*.

Baikalian polychaets of genus *Manayunkia* Leidy, 1859 (Sedentaria, Sabellidae) is not well-studied group. These organisms are representatives of one of few polychaets genera, living in fresh or brackish waters. The species, described by M.M. Kozhov (Kozhov, 1962) as *Manayunkia baicalensis*, was recently subdivided into three separate species on the basis of morphological and ecological features (Sitnikova *et al*., 1997). It's also confirmed with 18S rRNA data (Sitnikova *et al*., 1995). Baicalian polychaets also have relatively low mobility, and one to suppose, that vicariance can play the important role in diversification of these animals. It means, that intraspecies genetic distances in the given group of organisms should be described by model of "Isolation by Distance" (Wright, 1943).

EXPERIMENTAL DATA

In order to study intra-populationary polymorphism, we took the species *M. godlewskii*. This species inhabits generally on the silting bottom and has body size slightly more then two other ones. The animals were collected during the expeditions and fixed with 70 %

EtOH, or delivered alive. The DNA extraction was carried out with standard phenol chloroform method (Sambrook *et al*., 1967). As a marker for molecular phylogenetic researches we choose Folmer's fragment of 1 subunit of cytochrom-*c*-oxidase gene (mtCO1) (Folmer *et al*., 1994). The sequences obtained were aligned with using the program Bio Edit, version 5.0.9. All data obtained were classified according to location into four groups: 1) cape Berezovyi on south-western coast of Baikal, 2) Varnachka, 3) bay Peschanaya, and 4) Sahurta (strait Olkhonskie Vorota). These groups, in turn, are clustered into two groups: #1 (Berezovyi-Varnachka), and #2 (Peschanaya-Sahurta).

The analysis of genetic polymorphism was carried out by using of criterion F_{st} for DNA sequence data (Hudson *et al*., 1992). Values of the criterion vary in range from 0 up to 1. Small values of F_{st} correspond to higher gene flow between populations. The dependence of the criterion value on the physical distance between samples was obtained. For almost immobile organisms one may expect rapid growth of the values with the increase of distance between localities. But in our case it was fount that the distance between localities and criterion value are not correlated (Table 1).

Table 1. F_{st} criterion in real population and different theoretical model

F_{st} in real population (*M. godlewskii)*		F_{st} in model population (one species)		F_{st} in model population (two species)	
Distance (relative unit)	F_{st}	Distance (relative unit)	F_{st}	Distance (relative unit)	F_{st}
0.08	0.25	0.08	0.1	0.08	0.25
0.33	0.2	0.33	0.15	0.33	0.2
0.41	0.13	0.41	0.3	0.41	0.25
0.63	0.52	0.63	0.6	0.63	0.27
0.94	0.22	0.94	0.7	0.94	0.18
1	0.3	1	0.75	1	0.23

MODELS OF POPULATION DYNAMICS

Methods of population dynamics were applied in order to model the accumulation of genetic difference in highly viscous population(s). The model is based on the following assumptions: 1) all organisms do not move except for offspring at early stages, 2) the probability of fertilization of a female by male depends on distance between them and drops with the increase of distance, 3) the probability of death of each organism depends on a competition with other organisms, intensity of a competition depends on the density of organisms. All these assumptions follow the characteristics of the organisms studied. The general of the equations population dynamics thus will look as follows:

$$\frac{\partial N(x,t)}{\partial t} = rN(x,t)\left(1 - \frac{\int_{y_{\min}}^{y_{\max}} C(x-y)N(y)dy}{K}\right) - D\frac{\partial^2 N(x,t)}{\partial x^2} \qquad (1)$$

In this equation $N(x,t)$ there is a number of organisms in a point of space with coordinate x, r there is a speed of reproduction, K there is a capacity of an environment, D the factor of diffusion determining speed of distribution of organisms. Function $C(x-y)$ defines dependence of intensity of a competition on remoteness of organisms in space, in our model $C(x-y) = \frac{1}{2\pi\sqrt{\sigma}}\exp\left(-\frac{(x-y)^2}{2\sigma_C^2}\right)$.

Being based the equation (1), individually basев model describing population dynamics of investigated organisms has been developed. The probability of death on each time step is defined by the following ratio:

$$P_x = r \frac{\sum_{i=1}^{N} C(x - y_i)}{K} dt \qquad (2)$$

In this equation P_x is the probability of death of an organism with coordinate x, dt is the interval of integration. The probability of fertilization of the female by a given male depends on the distance between them according to Gaussian function $C_r(x_m - x_f, \sigma_r)$ where x_j and x_m there are coordinates of the female and the male. Thus σ_r equals σ_c. It means that sufficiently distant organisms don't interact. Distribution of the offspring is believed to follow the Gauss function $C_r(x_j,\sigma_D)$ where σ_D is proportional diffusion factor. Each organism in our model has neutral nucleotide sequence with length 500 base pairs long. Sequence is transferred from female to all offspring thus following neutral molecular evolution in mitochondrial DNA.

The numerical experiments, following initial and boundary conditions were used: the one-dimensional distribution range was represented by stretch l unit long with coordinates ranging from 0 to 1. At a simulation start the organisms had coordinates close to 0 and all neutral sequences were supposed to be identical. Periodically during calculation genetic sequences of the organisms living at edges of the range were used for calculation of F_{st}. Calculation was carried out until then while F_{st} did not achieve the stationary value (Table 1).

For each set of the nucleotide sequences resulting a simulation F_{st} criterion was calculated. Dependence between F_{st} on the distance between groups of organisms was found to be nonlinear, for closely located groups of organisms. F_{st} monotonously increases with the distance until it reaches saturation. This result correlates with the model "Isolation by Distance" but contradicts the real data from Baikalian polychaets.

Then the model is changed so that one-dimensional area is inhabited by 2 species of competing organisms which do not cross-breed. The competition between the organisms belonging to different species was set to be slightly less then between the conspecific ones. It was next scenarios of the mechanism of formation of polymorphism in neutral DNA sequences. The equations of population dynamics in this case will be:

$$\frac{\partial N_1(x,t)}{\partial t} = r_1 N_1(x,t) \left(1 - \frac{\int_{y_{\min}}^{y_{\max}} C(x-y) N_1(y) dy}{K} - \alpha \int_{y_{\min}}^{y_{\max}} C'(x-y) N_2(y) dy \right) - D \frac{\partial^2 N_1(x,t)}{\partial x^2} \qquad (3)$$

$$\frac{\partial N_2(x,t)}{\partial t} = r_1 N_2(x,t) \left(1 - \frac{\int_{y_{\min}}^{y_{\max}} C(x-y) N_2(y) dy}{K} - \beta \int_{y_{\min}}^{y_{\max}} C'(x-y) N_1(y) dy \right) - D \frac{\partial^2 N_2(x,t)}{\partial x^2}$$

In this equation $N_1(x,t)$ and $N_2(x,t)$ is number of organisms of competing species.

Value of F_{st} criterion was calculated as previous case, we suppose that information about intraspecies differences is unknown. In this model the value of F_{st} criterion does not depend on the distance between groups of organisms. The cause of it is next. Two species accumulate intraspecific differences connected with remoteness in space and interspecific differences connected with interspecific isolation. If it is not possible to distinguish the given organisms and they will be determined as one species in this case among intraspecies genetic distances there are interspecies genetic distances. It will bring in distortions at calculation F_{st}.

RESULTS AND DISCUSSION

Summarizing the results obtained it is possible to conclude, that the *Manayunkia* behaves according to the second model which assumes two species. Therefore the pattern of genetic diversity observed may be explained by assumption that *M.godlewskii* in fact consists of at least two sibling species which do not cross-breed.

REFERENCES

Folmer O., Black M., Hoeh W., Lutz, Vrijenkoek R. (1994) DNA primers for amplification of mitochondrial cytochrome C subunit I from diverse metazoan invertebrates. *Mol. Mar. Biol. Biotech*., **4**, 294–299.

Hudson R.R., Slatkin M., Maddison W.P. (1992) Estimation of levels of gene flow from DNA sequence data. *Genetics*, **132**, 583–589.

Kozhov M.M. (1962) Biology of Lake Baikal. M.: Nauka, 315. (In Russ.).

Sambrook J., Foitish E.F., Maniatis T. (1967) *Molecular cloning: A laboratory manual*. Goldspring Harbor Lab. Press, New York.

Sitnikova T.Ya., Sherbakov D.Yu., Ogarkov O.B., Sherbakova T.A. (1995) On the phylogenetic interrelashionships of Baikalian polychaets. *The Second Vereshchagin Baikal Conference, Irkutsk*, 177–178. (In Russ.).

Sitnikova T.Ya., Sherbakov D.Yu., Kharchenko V.V. (1997) On taxonomic status of polychaets of the genus *Manayunkia* (*Sabellidae, Fabricinae*) from the open Lake Baikal. *J. of Zoology*, **1**(76), 16–27. (In Russ.).

Wright S. (1943) Isolation by distance. *Genetics*, **28**, 114–138.

SEARCHING FOR AGROBACTERIAL T-DNA FRAGMENTS IN PLANT GENOMES

Chumakov M.I.**, *Mazilov S.I., Zotova T.V.
Institute of Biochemistry and Physiology Plants and Microorganisms, RAS, Saratov, 410049, Russia
* Corresponding author: e-mail: chumakov@ibppm.sgu.ru

Key words: *Agrobacterium*, T-DNA, horizontal transfer, corn, tobacco, petunia, trefoil, Arabidopsis, evolution

SUMMARY

Motivation: The aim of this work was to search for the nucleotide sequences in plant genome data banks similar to agrobacterial T-DNA fragments, in order to evaluate the role of naturally associated soilborne agrobacteria in plant evolution.

Results: Depending on the variant and length of the T-DNA right-border fragment, we found from 2 to 115 nucleotide sequences within different plant genomes. Most of them were found in the corn genome. The length of T-DNA right border-like sequences varied from 10 to 17 bp, depending on the plant genome. We hypothesize that the full-length agrobacterial T-DNA insertion into the plant genome possible involved in plant evolution but was eliminated during plant evolution, since T-DNA fragments are presented as 40–60 % of full-length TRBLF. All fragments similar with nptII and rolC genes found in plant genomes were very short (10–21 nucleotides) and possibly were not represented as full-length transfer sequences.

INTRODUCTION

Members of the genus Agrobacterium (family Rhizobeaceae) are natural soilborne plant-root-system residents that can transfer a portion of their Ti-plasmid DNA (T-DNA) into host-plant nucleus under condition of virulence-gene activation. In the presence of the VirD1 protein, VirD2 cleaves the border sequence in a site- and strand-specific manner and concomitantly becomes covalently attached to the 5'-end of the nicked T-DNA. *A. tumefaciens* transfers the ssT-DNA-VirD2 complex to the plant nucleus, where it becomes integrated in the plant chromosome, by using VirD2 and the plant repair system proteins in a sequence-independent manner (Chumakov, 2001).

We assumed that T-DNA might serve as a mutation factor to improve plant adaptation to the environmental conditions. The aim of this work was to search for nucleotide sequences similar to agrobacterial T-DNA fragments in plant-genome data banks and to evaluate the role of naturally associated soilborne agrobacteria in plant evolution.

METHODS AND ALGORITHMS

For computer searching for nucleotide sequences (GGCAGGATATT(CA/GG)G(T/G) TCTAA(AT/TC)) from agrobacterial T-DNA right border (Armitage *et al*., 1988), the genes nptII, rolC described in Chen *et al*. (2003) in plant-genome sequence databases (GenBank, DDBJ – DNA Data Bank of Japan) we used the BLAST program 2.2.14, and 2.2.12 versions)

at http://www.ncbi.nlm.nih.gov and http://www.ddbj.nig.ac.jp/ search /blast-e.html, respectively, and Clustal X 1.81 program (Higgins *et al*., 1996) for alignment of the sequences. All the checked variants of T-DNA right borders are listed in Table 1.

Table 1. Description of oligonucleotides from the T-DNA right border for *in silico* studies

1) ggcaggatattcagttctaaat; 2) ggcaggatattggggtctaatc; 3) ggcaggatattcagttctaatc;
4) ggcaggatattcaggtctaaat; 5) ggcaggatattcaggtctaatc; 6) ggcaggatattggggtctaaat;
7) ggcaggatattgggttctaaat; 8) ggcaggatattgggttctaatc

IMPLEMENTATION AND RESULTS

We found from 2 to 115 nucleotide sequences similar to the T-DNA right border-like fragments (TRBLF) in different plant genomes, depending on the variant and length of the TRBLF (Table 2). Most of the TRBLFs were found in the corn genome. The length of the TRBLF fragments found in the corn and arabidobsis genome ranged from 10 to 17 bp.

Table 2. The total number of T-DNA right border fragments** observed in the plant genomes

T-DNA right border-like fragments*	*Zea mays*****	*Petunia* sp. E<=214***	*Nicotiana tabacum* E<=437***	*Trifolium repens* E<=7.5***
1	115	21	17	9
2	83	16	20	4
3	96	21	20	10
4	62	19	27	3
5	80	20	29	3
6	95	16	15	2
7	93	9	22	2
8	95	9	20	2

* According to Table 1; ** The sum of 10–15 nucleotide fragments; *** the Expect value (E) is a parameter that describes the number of hits one can "expect" to see just by chance when searching a database of a particular size. It decreases exponentially with the Score (S) that is assigned to a match between two sequences. Essentially, the E-value describes the random background noise that exists for matches between sequences. For example, an E-value of 1 assigned to a hit can be interpreted as meaning that in a database of the current size one might expect to see 1 match with a similar score simply by chance. This means that the lower the E-value, or the closer it is to "0" the more "significant" the match is. However, keep in mind that searches with short sequences, can be virtually indentical and have relatively high E-value. This is because the calculation of the E-value also takes into account the length of the Query sequence. This is because shorter sequences have a high probability of occurring in the database purely by chance. For more details please see the calculations in the BLAST Course (http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html). The default value (10) means that 10 such matches are expected to be found merely by chance; **** 15 nucleotides (E = 1,8), 14 nucleotides – (E = 7,0), a 13 nucleotides (E = 28).

The number of the T-DNA right border-like oligonucleotides fully coinciding with the corn genome sequences is listed in Table 3.

Table 3. The number of the TRBLF fully coinciding with the corn genome sequences

T-DNA right border-like fragments*	The length of plant sequences of fully coincided with T-DNA right border fragments (bp)				
	15	13	12	11	10
1	1	5	31	30	48
2	–	–	11	44	28
3	1	2	30	29	35
4	–	2	10	37	13
5	–	2	9	41	28
6	–	1	14	46	34
7	–	6	13	43	31
8	–	3	12	42	38

* See Table 1.

We tried to search for the gene *npt*II in different plant genomes and found it (with an identity of 99.9 % and E = 0) only in the *eIF-4A1* gene for the translation initiation factor eIF-4A1 (exons 1–5 from the *Arabidopsis thaliana* plant) and within the *gus* gene for β-glucuronidase protein, as a result of transfer of binary vector pBI121 to the *Arabidobsis* genome (Chen *et al*., 2003). The 18-nucleotide-long *npt*II fragment (cgacggcgatgatctcgt) was also found in the *Arabidobsis* genome.

Within the *Arabidobsis* genome, we found a set of 18-nucleotide-long fragments (E = 1.8) from the gene *rol*C. Within the *Zea mays* genome, we found a set of 21-nucleotide-long fragments (E = 0.042); a set of 16-nucleotide-long fragments (E = 11) was found within the *Tobacco* genome originating from the gene *rol*C.

Earlier, Matveeva *et al*. (2004) by using hybridization experiments did not find *rol*C in the *Arabidobsis* genome.

Thus, by using BLAST program, we found a set of short and presented by: 10–15 nucleotides or 40–60 % of full-length TRBLF) T-DNA right border-like fragments, and *npt*II-like (0.01–0.02 % of full-length *npt*II) fragments within different plant genomes and data banks; and 18–21 nucleotides fragments from the gene *rol*C (0.03–0.04 % of full-length gene) in the corn, *Arabidobsis* and *Tobacco* genomes.

We hypothesize that the full-length agrobacterial T-DNA insertion into the plant genome possible involved in plant evolution was eliminated during plant evolution, since T-DNA fragments are presented as 40–60 % of full-length TRBLF. All fragments similar with *npt*II and *rol*C genes found in plant genomes were very short (10–21 nucleotides) and possibly were not represented as full-length transfer sequences.

REFERENCES

Armitage P. *et al*. (1988) Agrobacterial vectors for plant transformation In Draper J., Armitage P., Scott R. (eds), *Plant Genetic Transformation and Genetic Expression. A Laboratory Manual.* Blackwell Scientific Publishing. pp. 11–86.

Chen P.Y. *et al*. (2003) Complete sequence of the binary vector pBI121 and its application in cloning T-DNA insertion from transgenic plants. *Mol. Breed.*, **11**, 287–293.

Chumakov M.I. (2001) *The mechanism of agrobacterial transformation of plants*. Saratov, Slovo. 256 p.

Higgins D.G. *et al*. (1996) Using CLUSTAL for multiple sequence alignments. *Methods Enzymol.*, **266**, 383–402.

VARIATIONS IN NUCLEOTIDE COMPOSITION OF THE REGION ITS1-5.8S RDNA-ITS2 IN EVOLUTIONARY ADVANCED AND EVOLUTIONARY STATIC BRANCHES OF THE MONOCOTYLEDONOUS PLANTS

Chupov V.S.**, *Machs E.M.
Komarov Botanical Institute, RAS, St.Petersburg, Russia
* Corresponding author: e-mail: nika-egida@mail.ru

Key words: ITS1, CpG, dinucleotides, Melanthiales, Asparagales, evolution

SUMMARY

It was demonstrated the correlation between the CpG contents, dinucleotide spectrums, the length and contents of oligonucleotide tracks and peculiarities of evolution in phylogenetic lines *Melanthiaceae – Liliaceae* and *Melanthiaceae – Asparagaceae*. Authors interpret differences of nucleotide composition in two phylogenetic lines of *Monocotyledonous* as the indication of considerable changes of DNA replication and reparation systems during saltation periods of plant evolution.

INTRODUCTION

Evolutionary phenomenon of decrease of CpG and increase of TpG and CpA content was reported by many authors (Mazin, Vanyushin, 1987a; Frixel, Zukerkandl, 2000; Vanyushin 2005). Different degree of methylation of CpG in different genes (Mazin, Vanyushin, 1987b) and following substitution of 5-methylcytosine by thymine appears to be a mechanism of light and heavy isochores origin. Afterwards the methylation of other sites was shown (Volpe, 2005). At present time many authors consider the conversion of 5-methylcytosine into thymine as a key process of genome evolutionary changes.

On the other hand we demonstrated considerable increase of cytosine, guanine and CpG contents in ascending phylogenetic lines of *Monocothyledons* (Chupov *et al*., 2003, 2004, 2006). The study of the evolution of amino acid composition of proteins demonstrated the increase of prolyne, alanyne, glutamine and glycine and decrease of cysteine, methyonine, gysthydine, serine and phenylalanine contents in a wide range of taxa (Bazykin *et al*., 2004; Jordan *et al*., 2005). These obsevations also are against the concept of decrease of cytosine and increase of adenine and thymine in evolutionary process because of the prolyne, alanyne, and glycine codons are GC-enriched. Therefore the problem of the study of different types of mutagenesis in evolutionary transformation of genomes appears to be very important.

MATERIAL AND METHODS

We analyzed sequences available in the NCBI GeneBank database including several samples of *Paris* and *Trillium* sequenced by authors. We used the software DAMBE for the analysis of dinucleotide spectrums and Vector NTI for the analysis of oligonucleotide tracks.

RESULTS AND DISCUSSION

We studied the nucleotide and dinucleotide contents of ITS1, 5.8S rRNA and ITS2 regions of several Monocotyledonous taxa pertaining to phylogenetic lines *Melanthiaceae – Liliaceae* and *Melanthiaceae – Asparagaceae*. Among all studied regions the highest correlation between morphological characters and nucleotide composition was shown for ITS1 region. It was demonstrated that the contents of CpG in evolutionary advanced groups is considerable higher as compared with primitive ones (Chupov *et al*., 2003, 2004, 2006). The CpG contents in ITS1 per 100 nucleotides for taxa arranged by the level of evolutionary advance from primitive Melanthiales to high evolutionary advanced Asparagales is shown on Fig. 1. Dinucleotide spectrums also change (Fig. 2, 3).

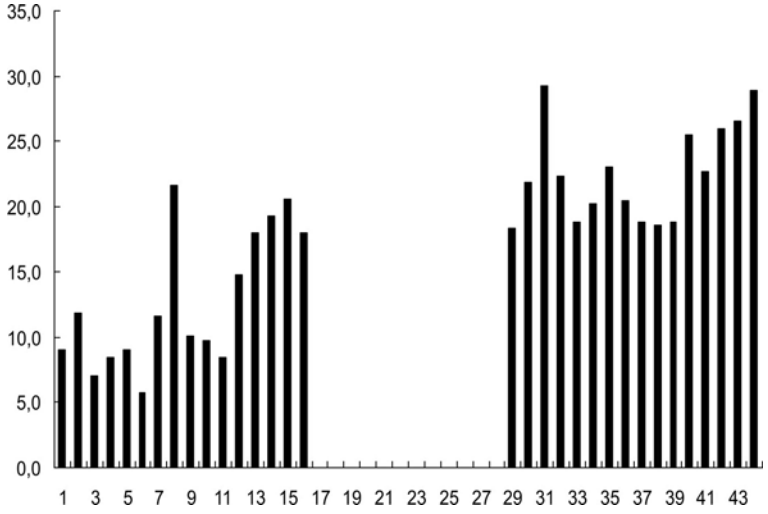


Figure 1. CpG + CpNpG contents in phylogenetic line Melanthiales-Asparagales.

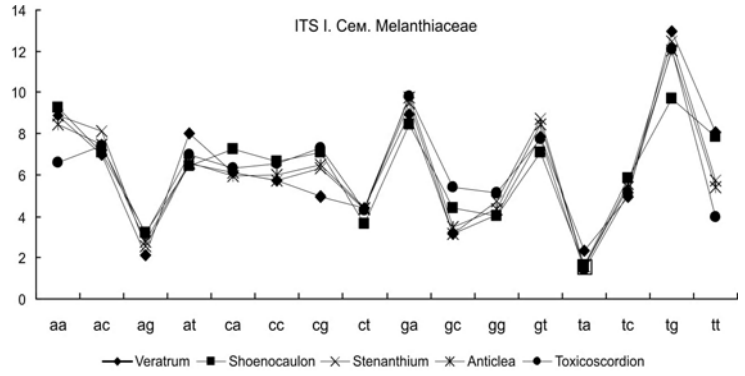


Figure 2. Dinucleotide spectrums of ITS1 of evolutionary primitive *Monocothyledonous* (*Melanthiaceae*).

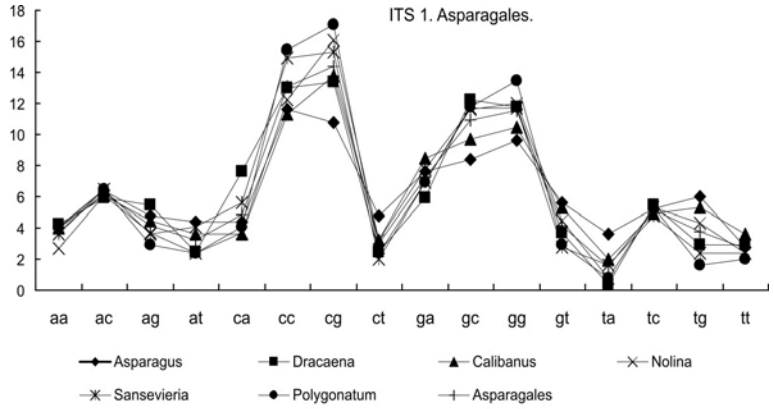


Figure 3. Dinucleotide spectrums of ITS1 of evolutionary advanced *Monocothyledonous* (*Asparagales*). 1 – *Anticlea*, 2 – *Toxicoscordion*, 3 – *Veratrum*, 4 – *Melanthium*, 5 – *Stenanthium*, 6 – *Amianthium*, 7 – *Shenocaulon*, 8 – *Zigadenus*, 9 – *Kinugasa*, 10 – *Daiswa*, 11 – *Paris*, 12 – *Trillium*, 13 – *Cardiocrinum*, 14 – *Lilium*, 15 – *Nomocharis*, 16 – *Gagea*, 29 – *Bulbine*, 30 – *Aloe*, 31 – *Hosta*, 32 – *Camassia*, 33 – *Beschorneria*, 34 – *Manfreda*, 35 – *Furcraea*, 36 – *Agave*, 37 – *Hesperaloe*, 38 – *Yucca*, 39 – *Asparagus*, 40 – *Dracaena*, 41 – *Calibanus*, 42 – *Nolina*, 43 – *Sansevieria*, 44 – *Polygonatum*. The number of dinucleotides per 100 nucleotides of the length of ITS1 is shown on Y axis.

The analysis of aligned sequences demonstrated the presence of not only CpG → TpG + CpA mutations but also considerable changes of homonucleotide tracks (CCCC, GGGG and others). Because of the ambiguity of the alignment of sequences of distant taxa we used statistical analysis of homonucleotide tracks in studied phylogenetic lines (Table 1). These data demonstrates the decrease of length and contents of oligo-A and oligo-T tracks and the increase of corresponding values for oligo-C and oligo-G tracks with the increase of the evolutionary advance level (the level of evolutionary advance increase from top to bottom in the Table).

This observation probably indicates the role of replication and reparation mechanisms in DNA evolutionary changes while due to its random or regular "errors" the length of homonucleotide sequences can change.

The concept of chemical modification of nucleotides can not solve the problem of evolutionary increase of CpG value. Thus cytosine–thymine conversion is easy but the reverse process is impossible. Partially this problem can be interpreted in the following way. The replication "errors" result in 5′-3′ extension of oligo-C sequences eliminating T (transitions) up to first G (transversions). The reparation "errors" result in oligo-G tracks extention in opposite direction up to first C. This assumption can be confirmed by the analysis of ITS1 sequences in phylogenetically related taxa. Thus in taxa representing saltaion groups (*Zigadenus*) 36 % of CpG dinucleotides are located before guanine(s) whereas in other groups of Melanthiales representing gradual evolution the contents of such dinucleotides is only 17 %. Another way of the increase of CpG contents is the splitting of C-tracks by the insertion of guanine. The problem of the absence of accumulation of CA and TG dinucleotides is not clear now. Next step of molecular evolution (*Trilliaceae)* destroy a lot of oligo-C and oligo-G tracks but the contents of C and G remain higher as compared with *Melanthiaceae.* The same changes we observe in transition to next evolutionary advanced groups *Liliaceae* and *Agavaceae*. In this case saltation group is represented by *Cardiocrinum giganteum* and *Hosta ventricosa*.

We demonstrated (Chupov *et al*., 2004, 2006) that changes of nucleotide spectrum not only in general but also in details related with evolutionary saltations correlate with evolutionary progress in plants. Prolonged vectorial changes of nucleotide contents equal in different phylogenetic lines highly coinciding with morphological structure is the evidence of complex molecular genetic mechanisms underlie the process of plant evolution.

Table 1. The number 2, 3, 4, 5, 6, 7 length oligo sequences and total number of nucleotides (n) in corresponding sequences in ITS1 region of rDNA of phylogenetic lines Melantiales – Liliales and Melantiales – Asparagales

Taxa	A						T					C							G						
	2	3	4	5	6	n	2	3	4	5	n	2	3	4	5	7	8	n	2	3	4	5	6	7	n
Melanthiales																									
Melanthiaceae																									
Anticlea virescens	18	2				42	9	2			24	8	1	2				27	5	3					19
Anticlea elegans	18	2				42	9	2			24	8	1	2				27	5	3					19
A. occidentale	17	2		1		45	11	2	1	1	29	6	1					15	5	1					13
Amianthium muscitoxicum	18	2				36	9	2			24	8	1	2				27	5	3					19
Melanthium latifolium	15			1		35	14	3			37	7	2					20	7	2					20
Toxicoscordion nuttaliana	11	1		1		30	6	2			18	12	1	1				31	7	2					20
T. paniculatus	11	1		1		30	6	2			18	11	1	1				29	7	3					23
Zigadenus glaberrimus	8		1		1	26	2	1			7	5		3	2			32	14	5	1				47
Trilliaceae																									
Kinugasa japonica	12	2	3			42	3	3			15	5	3					19	10	3					29
Daisva fargesii	14	3	1			41	7	3			23	3	3	1				19	10	3	1				33
D. poliphylla	13	3	1			39	7	3			23	2	2	2				18	12		1				28
Paris incompleta	11	2	1			32	3	2	1		16	7	1	1				21	9	2			1		30
P. verticellata	12	2	1			34	3	1	1		13	7	2	2				28	11	2		1			33
Trillium camtchatcense	13	3	1			39	3	1	1		13	4	3	1				21	10	3	1				33
T. chloropetalum	12	2	1			34	3	2			12	4	3	2				25	12	2		1			35
Liliales																									
Liliaceae																									
Cardiocrinum gigantheum	8		3			28		4			4	13	3					35	9	2	1	1			33
Lilium pyrenaicum	8	1	1			23	3	2			12	13		1				30	13	1	1	1			38
L. candidum	8	2	1			26	3	2			12	13						26	10	2		1			31
L. formosanum	10	2				26	7	1			17	9	2	1				28	12	4		1			41
L. henryi	6	3				21	2	2			10	16	1	1				39	12	2	1	1			39
Asaragales																									
Amaryllidaceae																									
Zephyranthes atamasco	10	1				23	5	1			11	9	3	1				31	11	4				1	41
Funkiaceae																									
Hosta ventricosa	7	1				17	5				10	10	8	1	1			53	14	5	2	2	1		67
Agavaceae																									
Agave americana	9	1				21	7	2			20	14	2		1			39	10	4	2	2			50
A. attenuata	9	1				21	7	2			20	15	2		1			41	10	4	2	2			50
Yucca treculeata	10	1				23	8	1			19	11	6		1			45	12	2	3	1			47
Y. whipplei	10					21	7	2			20	12	7		1			50	13	1	3	1			46
Asphodelaceae																									
Aloe acutissima	6	4				24	2	3			13	10	5	1			1	48	14	4					40
A. bainesii	7	2				20	3	2			12	10	4	2		1		48	12	5					39
A. cremnophylla	5	2	1			20	4	3			17	10	4	2		1		48	12	4					36
Poligonataceae																									
Polygonatum biflorum	6					12	2		1		8	12	3	4	1			54	12	6	1				46

ACKNOWLEDGEMENTS

The works was suppoted by the RFBR grant 06-04-48399.

The authors are indebted to stuff members of the Laboratory of Biosystematics and Plant Cytology Komarov Botanical Institute RAS for helpful discussions.

REFERENCES

Bazykin G.A., Kondrashov F.A., Ogurtsov A.Y., Sunyaev S., Kondrashov A.S. (2004) Positive selection at sites of multiple amino acid replaciments since rat-mouse divergence. *Nature,* **429**, 558–562.

Chupov V.S., Machs E.M., Rodionov A.V. (2006) Dinucleotide composition of the ribosomal spaicer regions ITS1-5.8S rDNA-ITS2 as a systematic character and a phylogenetic marker of macrotaxons of the monocotyledon plants: general direction of the dinucleotide value variations in the evolution of Melanthiaceae, Iridaceae, Trilliaceae и Liliaceae. *Tsitologia* (In press).

Chupov V.S., Punina E.O., Machs E.M., Rodionov A.V. (2003) The contents of CpG in ribosomal cluster. *Abstracts of VGO Congress,* **2**, 223–224. (In Russ.).

Chupov V.S., Punina E.O., Machs E.M., Rodionov A.V. (2004) Mutational process in transition taxa between *Melanthiaceae* and *Trilliaceae*. *Genetica in XXI century: current state and perspectives*. **2**, 500. M: Nauka. (In Russ.).

Frixel K., Zuckerkandl E. (2000). Cytosine deamination plays a primary role in the evolution of mammalian isochores. *Mol. Biol. Evol*., **17**, 1371–1383.

Jordan I., Kondrashov F.A., Adjubei I., Wolf Y., Koonin E.V. (2005) A universal trend of amino acid gain and loss in protein evolution. *Nature,* **433**, 633–638.

Mazin A.L. (1995) Methylation of the factor IX gene – a basic reason for the mutation causing hemophilia B. *Mol. Biol*. (Mosk), **29**, 71–90. (In Russ. The resume in English).

Mazin A.L., Vanyushin B.F. (1987a) The loss of dinucleotides CpG from DNA. IV. Methylation and divergence of genes and pseudogenes of small nuclear RNA. *Mol. Biol*. (Mosk), **21**, 1099–1110. (In Russ. The resume in English).

Mazin A.L., Vanyushin B.F. (1987b) The loss of CpC dinucleotides from DNA. II. Methylated and non-methylated genes of vertebrates. *Mol. Biol*., **21**, 552–562. (Mosk). (In Russ. The resume in English).

Vanyushin B.F. (2005) Ensyme methylation of DNA as an epigenetic control mechanism of cell genetic functions. *Biokhimiia* (Mosk), **70**, 598–611. (In Russ.).

Volpe P. (2005) The language of methylation in the genomics of *Eucariota*. *Biokhimiia* (Mosk), **70**, 708–721. (In Russ.).

HUMAN-CHIMPANZEE PROPERTY-DEPENDANT COMPARISONS ON CHROMOSOMES 21

***Deyneko I.V.*[*1, 2], *KalybaevaY.M.*[1], *Kel A.E.*[3], *Blöcker H.*[*1], *Kauer G.*[4]**
[1] Department of Genome Analysis, GBF, Braunschweig, Germany; [2] Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia; [3] BIOBASE GmbH, Wolfenbüttel, Germany; [4] University of Applied Sciences, Emden, Germany
[*] Corresponding authors: e-mail: ide@gbf.de; bloecker@gbf.de

Key words: comparative genomics, property-dependant similarities, FeatureScan

SUMMARY

Motivation: Identification of different functional elements and their properties is a fundamental need in biomedical research, and phylogenetic comparisons form a solid basis for this task. But being applied to close genomes they can not "feel" such small, but nevertheless phenotypically important differences in sequence.

Results: In this work we present the comparative analysis of two evolutionary close genomes – human and chimpanzee. In contrast to previous studies we focus on evolutionary differences and evaluate changes in DNA properties rather than count nucleotide mismatches. In our examples, we find that nucleotide mismatches in promoters were probably introduced in a correlated manner during the course of evolution. Such property-dependant conservation of promoters is significantly different from nucleotide conservation and shows significant functional biases.

INTRODUCTION

Comparative genomics provides a powerful approach for investigating newly sequenced genomes. To date the great advances in this area were achieved in the identification of either protein-coding or non-protein-coding functional elements (Xie *et al.*, 2005). The application of traditional techniques for comparisons of evolutionary close species (as for example, human and chimpanzee) is not effective, since a substantial amount of functional elements are masked by prolonged conserved non-functional DNA stretches.

Genomic sequences of human and chimpanzee are roughly 99 % conserved, but obvious differences in both, appearance and behaviour are surely beyond any doubt. Therefore, phylogenetic comparisons together with novel methods for long-range sequence comparison, which are able to "feel" small differences between genomes, have the potential to reveal certain key mechanisms in evolution.

In a first test case, we decided to compare promoters of genes located on chromosomes 21 from human and chimpanzee with our novel signal-theory-based approach to long range property-dependant sequence comparison (Deyneko *et al.*, 2005).

DATA AND METHODS

Using the *Ensembl* database (v. 37.3a) we built up two sets of upstream sequences of all orthologous genes on chromosome 21 from human and chimpanzee. Each sequence

spans 2Kb of the upstream region starting from the 5′-most annotated transcription start site. The total number of sequences in each set was 229.

Property-dependent similarities of orthologous promoters were calculated using our tool FeatureScan (Deyneko *et al.*, 2004) with "melting enthalpy" as DNA characteristic. Algorithmically, FeatureScan originates from proven methodologies in image analysis and speech recognition. The current implementation is based on a convolution method and can be described briefly in three main steps (for the detailed theoretical background see our earlier publication (Kauer, Blöcker, 2003)). First is a transformation of nucleotide sequences (pattern and investigated sequence) into numerical form, which we refer to as signals. At this step users have to decide which property may play an important role in their specific cases. Second is a computation of the correlation integral (i) of two signals $\boldsymbol{f}$ and $\boldsymbol{g}$, which can be rewritten using Fourier transformants $\boldsymbol{F}$ and $\boldsymbol{G}$ yielding (iii). Assuming to have direct and inverse Fourier transformations implemented (in our case it is hardware implemented), the entire integral is reduced just to a multiplication. The final step is looking for shift values $\boldsymbol{y}$ which will define possible matches of the sequences, so that the difference between correlation (iii) and autocorrelation (ii) integrals is less than the predefined threshold.

$$Corr(y) = \int f(x) \cdot g(x - y)dx \quad \text{(i)}$$

$$AutoCorr = \int g(x) \cdot g(x)dx \quad \text{(ii)}$$

$$Corr(y) = InverseFourierT\left\{F(y) \cdot \overline{G(y)}\right\} \quad \text{(iii)}$$

For statistical evaluation of the results we tried to simulate the evolution of primate promoters under pure random and transition/transversion biased models with transition/transversion rate ratio of 4.31 (Rosenberg *et al.*, 2003). The set of 10^4 random sequences were generated, which were of the same length (2Kb) and the same mononucleotide context as promoter regions. Calculation of *p-values* was done by summing the "tail" of the binomial distribution assuming the random value follows Bernoulli trials scheme.

RESULTS

Following the scheme described above, we calculated the similarities of chimpanzee and human promoter regions using letter-based (ClustalW) and signal-theoretical approaches (FeatureScan). The distribution of the number of promoters *vs*. similarity is shown in Fig. 1.

It can be easily seen that promoters of chimpanzee and human genes, which differ by less than 2 % of nucleotides, show significantly higher similarity by FeatureScan than expected (Fig. 1; white bar is higher than light grey and dark grey). If we consider the transition/transversion bias, than this increase will even be improved. We found that 139 out of 198 orthologous promoter pairs showed higher signal similarity than can be expected, which corresponds to a *p-value* of $2{,}43*10^{-8}$.

Using the EMBL-EBI gene ontology classification, we examined the gene distribution which showed high signal similarity of their promoters. A subset of 15 genes involved in the molecular function "metal ion binding" (GO:0046872) and another subset of 11 genes involved in "nucleotide binding" (GO:0000166) were identified. These observations have an estimated *p-value* of $4{,}9*10^{-3}$ and $4{,}52*10^{-5}$, respectively. Corresponding subsets of 3 and 4 genes, can be identified in the set of genes with low promoter similarity.

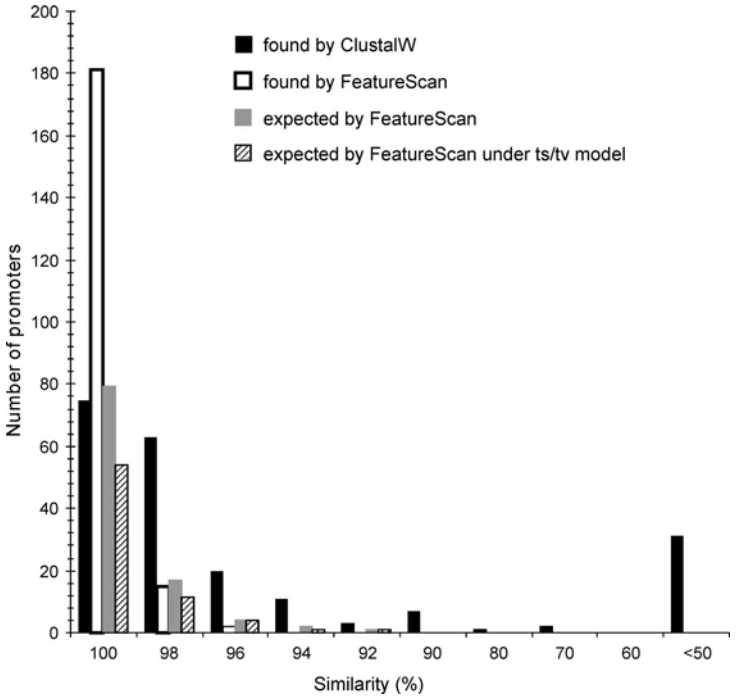


Figure 1. Number of orthologous promoters showing given level of similarity.

DISCUSSION

The main advantage of phylogeny of close species is the ability to "see" evolutionary tendencies, which are not yet "drowned" in mutation chaos. Here we investigated the similarity of 2Kb promoter regions of human and chimpanzee genes, which is based on melting enthalpy characteristic of DNA. The observed statistically significant overrepresentation of promoters with high property-dependant similarity encouraged us to speculate, that single mutations occurring in evolution tend to compensate disturbances involved by others to retain the "original" function. As we may conclude from the presented results, a nucleotide substitution decreasing the melting temperature of a locus, may induce evolutionary pressure for further changes in the close vicinity to outbalance the first mutation.

The promising advantages of property-dependant similarity measures encourage us to tackle further interesting problems. It might be interesting to investigate changes of characteristics (melting temperature, conformation and others) caused by SNP mutations across the entire human genome. We believe that SNPs (both in coding and promoter regions) which are correlated with diseases or have phenotypic evidences, should be detectable/distinguishable in a single- or multidimensional property space.

ACKNOWLEDGEMENTS

Financial support by the German Federal Ministry for Education and Research through Projektträger Jülich (FKZ 031U210A) is gratefully acknowledged.

REFERENCES

Chimpanzee Sequencing and Analysis Consortium. (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, **437**, 69–87.

Deyneko I.V., Kel A.E., Blöcker H., Kauer G. (2005) Signal-theoretical DNA similarity measure revealing unexpected similarities of *E. coli* promoters. *In Silico Biol.*, **5**, 547–555.

Deyneko I.V., Kel A.E., Wingender E., Gössling F., Blöcker H., Kauer G. (2004) Signal theory – an alternative perspective of pattern similarity search. *Proceedings of the Fourth International Conference on Bioinformatics of Genome Regulation and Structure*, Novosibirsk, Russia, **2**, 25–28.

Kauer G., Blöcker H. (2003) Applying signal theory to the analysis of biomolecules. *Bioinformatics*, **19**, 2016–2021.

Rosenberg MS, Subramanian S, Kumar S. (2003) Patterns of transitional mutation biases within and among mammalian genomes. *Mol. Biol. Evol.*, **20**, 988–93.

Xie X., Lu J, Kulbokas E.J., Golub T.R., Mootha V., Lindblad-Toh K., Lander E.S., Kellis M. **(**2005**)** Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature***, 434,** 338–45.

EVOLUTION AND ORIGIN OF NEUROFIBROMIN, THE PRODUCT OF THE *NEUROFIBROMATOSIS TYPE 1 (NF1)* TUMOR-SUPRESSOR GENE

***Golovnina K.*[*1], *Blinov A.*[1], *Chang L.-S.*[2]**
[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia; [2] Center for Childhood Cancer, Children's Research Institute, Children's Hospital and Department of Pediatrics, The Ohio State University, USA
[*] Corresponding author: e-mail: ksu@bionet.nsc.ru

Key words: neurofibromin, the *Neurofibromatosis type 1 (NF1)* gene, RasGAP protein family, GTPase-activator proteins for Ras-like GTPase, phylogeny, BLAST

SUMMARY

Motivation: Neurofibromatosis type 1 (NF1) is a common genetic disorder, which predisposes affected individuals to a variety of clinical features including tumors of the central and peripheral nervous systems. The product of the NF1 gene, neurofibromin, is a tumor suppressor which most likely acts through the interaction of its GTPase activating protein (GAP) related domain (GRD) with RAS to regulate cellular growth. While clinical features of NF1 as well as functional activity of human neurofibromin are intensively studied now little is understood about its evolution, diversity, and overall distribution among different taxa.

Results: By combining bioinformatic and phylogenetic approaches, we demonstrated that NF1 homologs are present across a wide range of eukaryotic lineages. We observed 26 similar to NF1 amino-acid sequences from Chordata, Echinodermata, Arthropoda, Platyhelmintes and Fungi taxons. Taking into account a presence of *NF1* gene in fungi, we can suggest the derivation of this gene before the Metazoan origin. In this case, an absence of *NF1* in Nematoda and Mollusca should be a result of this gene elimination.

INTRODUCTION

Clinical features of NF1. Neurofibromatosis type I (NF1) or von Recklinghausen neurofibromatosis is the most common cancer predisposition syndrome affecting the nervous system with the incidence of 1 in 3000 worldwide (Gutmann, 2001). Typical manifestations include café au lait spots (hyperpigmented macules), cutaneous and subcutaneous neurofibromas (benign tumors), and malignancies of the central and peripheral nervous systems. The less common abnormalities observed in NF1 patients include learning disabilities (although frank mental retardation is rare) and skeletal abnormalities such as scoliosisand pseudoarthrosis (Skuse, Cappione, 1997). In addition to these features, children with NF1 can present within the first 6 years of life with low-grade glial tumors involving the optic parthway. Two intriguing features of NF1 are the wide range of potentially affected tissues and the great variation in expressivity of disease traits across those affected. To date, the underlying source of this variation remains somewhat unclear, but evidence suggests that aberrations in normal *NF1* RNA processing may be involved (Skuse, Cappione, 1997).

Genetic features of NF1. Early insights into the patogenesis of NF1-associated tumors began with the identification of the *NF1* gene in 1990 (Wallace *et al*., 1990). The *NF1* gene itself, located at 17q11.2, encompasses >300 kb of human chromosome 17 (Li *et al.*, 1995). The 60 exons which constitute the human *NF1* gene give rise to several alternatively spliced transcripts. Within the central portion of the *NF1* encoded protein, neurofibromin, lies a region with homology with the mammalian GTPase activating proteins (GAP) and the yeast inhibitor of RAS proteins 1 and 2 termed the GAP related domain (GRD).

Neurofibromatoses. Discussions about neurofibromatoses include two the most common forms of this disorder: neurofibromatosis 1(NF1) and neurofibromatosis 2 (NF2). Individuals with two inherited cancer syndromes, NF1 and NF2 develop both benign and malignant tumors. The corresponding genes mutated in these two disorders encode tumor suppressor proteins, termed neurofibromin (NF1) and merlin (NF2), which have a similar function to regulate cell growth and differentiation. While both genes are not related and located on the different chromosomes in human, nothing is known about their relationships, distribution and function in other organism as well as about their origin.

In the previous study we investigated evolution of merlin and postulated its origin in early metazoan (Golovnina *et al.*, 2005). Here we represent the initial phylogenetic research of NF1 protein and its homologs based on bioinformatics approaches.

METHODS AND ALGORITHMS

BLAST search. Initial sequences of genes and proteins of interest from various organisms were identified by performing multiple TBLASTN and BLASTP (Altschul *et al*., 1997) searches against GenBank (http://www.ncbi.nlm.nih.gov/Genomes/), Ensembl (http://www.ensembl.org) and wormbase (www.worbase.org) databases. In each case putative *H.sapiens* NF1 protein were used as the query sequence. Only NF1-like representatives of insects were located by *D. melanogaster* NF1 protein homology. To obtain more information we then searched the desirable sequences across genomic databases of completely or partially sequenced genomes available at The Sanger Institute (http://www.sanger.ac.uk/DataSearch) and The Institute for Genomic Research (TIGR) (http://tigrblast.tigr.org/tgi/), Doe Joint Genome Institute (JGI) (http://www.jgi.doe.gov), The Broad Institute (http://www.broad.mit.edu). The predicted nucleotide and aminoacid sequences of many species were assembled manually using available contigs and assemblies of genomes by homology to query sequence.

Alignments and phylogeny. The Clustal X program (Thompson *et al.*, 1997) was used to align the characterized or predicted protein sequences from different species. All alignments were corrected for obvious alignment ambiguity The resultant alignment contained 3888 aligned positions and was used to construct phylogenetic tree. Phylogenetic analysis was carried out using the Neighbor-Joining method in MEGA 3.1 program (Kumar *et al.*, 2004).

RESULTS AND DISCUSSION

In total, 22 NF1-like sequences, that have overall homology, from Deuterostomia, Arthropoda, Platyhelmintes have been located in the present investigation. The main criterion for identification these sequences was their evident homology to known NF1 proteins. Moreover, we have found four Fungi proteins that show not so strongly similarity to NF1-like sequences. Five of 22 NF1-like proteins are experimentally annotated, namely *H. sapiens* NF1, *M. musculus* NF1, *R. norvegicus* NF1, *T. rubripes* NF1, and *D. melanogaster* NF1. Eight homologs were bioinformatics predicted previously and were located by initial BLASTP search across NCBI database

(*P. troglodytes*, *C. familiaris*, *G. gallus*, *A. gambiae*), as well as NF1-related protein of *N. crassa* and three other RasGAPs of fungi.

The rest sequences were assembled manually by parts using available paired scaffolds, contigs, and assemblies of the sequenced genomes together with protein and EST databases. Based on the obtained NF1 sequences and the close related proteins the phylogenetic tree was constructed (Fig. 1). RasGap family members from another groups were included in our analysis as an outgroup to show a homology of all obtained NF1-like sequences. We conducted phylogenetic analysis including only RasGap domain and corresponding sequences of newly identified proteins in view of possible different origin of some other domains except RasGap (Fig. 2) Both trees are similar in topology with small alterations in Fungi group.

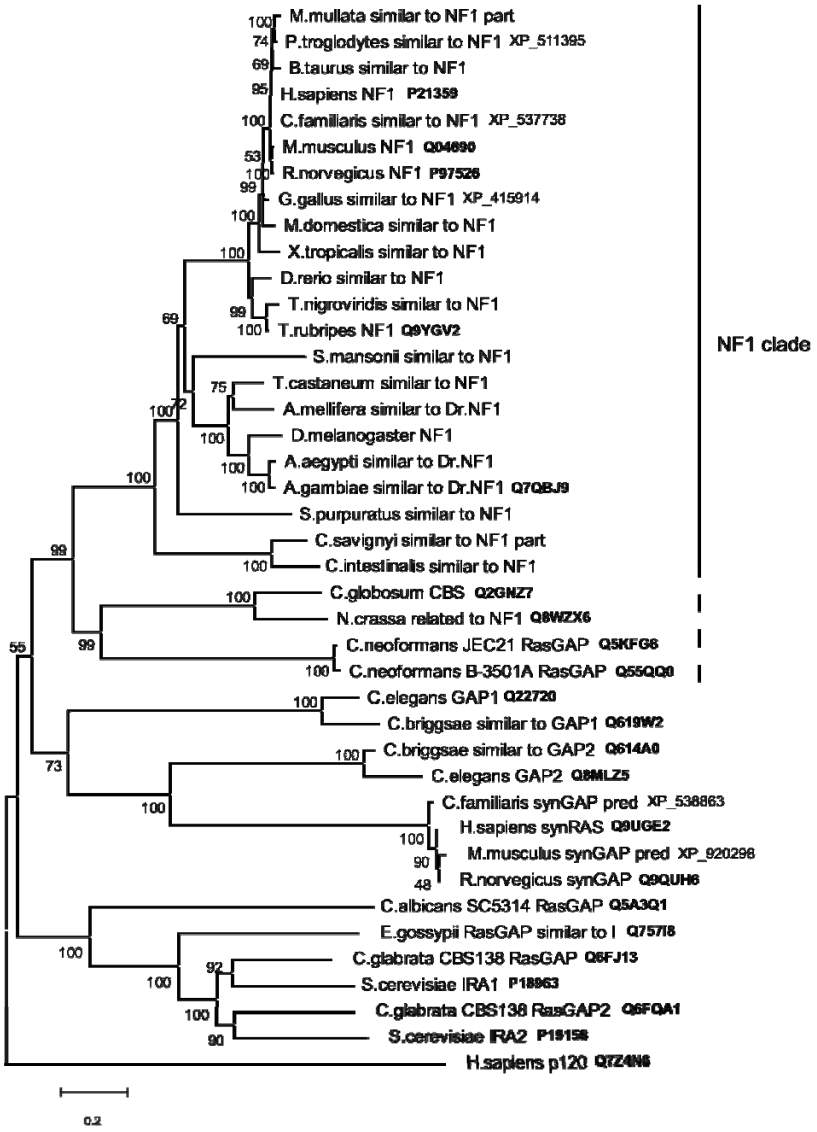


Figure 1. Phylogenetic analysis of NF1 homologs and its close related proteins from RasGAP group. NF1 clade is shown by solid line on the right. The dashed line denotes observed fungi proteins that are similar to NF1. UniProt accession numbers for full length sequences are represented in bold letters, GenBank – in regular.

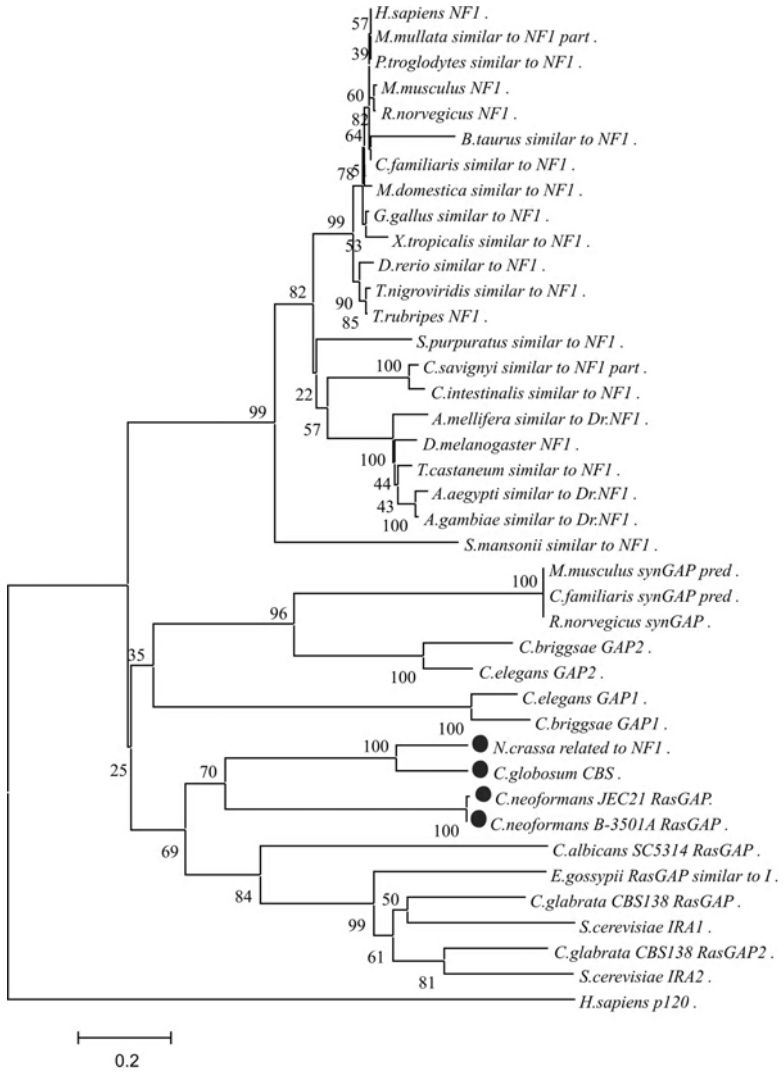


Figure 2. Phylogenetic analysis of NF1 homologs and its close related proteins from RasGAP group based on RasGap domain sequences. Fungi proteins that have similarity with NF1-like sequences are marked with dark round.

Phylogenetic analysis demonstrate a distribution of NF1 homologs across all investigated and available to date the major metazoan (chordata, urochordata, echinodermata, insects, and platyhelmintes) and fungi lineages. Among all species investigated, no NF1 sequences have been found in round worms (*C. elegans*, *C. briggsae*, *C. remanie*, *B. malayi*) and mollusk (*B. glabrata*). There are two possible explanations: (i) the derivation of *NF1* gene occured after a separation of both Nematoda and mollusca or (ii) it was lost specifically in these groups. In *C. elegans* genome were previously annotated RasGAP family members (GAP1 and GAP2) that have homologs in *C. briggsae* genome. They are clasterized together with synRAS group of proteins that means their evolutionary relatedness. The presence of NF1-related sequences in fungi is disputable. NF1-like proteins have been found in four species of fungi kingdom, belonging to both Ascomycota and Basidiomycota. These amino-acid sequences have overall similarity to other NF1-like proteins but not so strong besides sequences corresponding to RasGap domain are more similar to other fungi RasGap proteins (Fig. 2). One of them (*N. crassa*) was previously annotated as NF1-related protein. On the other hands, no NF1-like sequences were detected in the rest fungi species examined.

Therefore, the root of neurofibromin origin is unclear now and with the great progress in genome sequencing projects it promised to be more definite. Previously, no merlin homologs (another tumor suppressor) in fungi were observed.

ACKNOWLEDGEMENTS

This study was supported by grants from the US Department of Defense Neurofibromatosis Research Program.

REFERENCES

Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W., Lipman D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res*., **25**, 3389–3402.

Golovnina K., Blinov A., Akhmametyeva E.M., Omelyanchuk L.V., Chang L.-S. (2005) Evolution and origin of merlin, the product of the neurofibromatosis type 2 (NF2) tumor-suppressor gene. *BMC Evolutionary Biology*, **5**, 69.

Gutmann D.H. (2001) The neurofibromatosis: when less is more. *Hum. Mol. Genet*., **10**, 747–755.

Kumar S., Tamura K., Nei M. (2004) MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. *Briefings in Bioinformatics*, **5**, 150–163.

Li Y., O'Connell P., Breidenbach H.H., Cawthon R., Stevens J., Xu G., Neil S., Robertson M., White R., Viskochil D. (1995) Genomic organization of the neurofibromatosis 1 gene (NF1). *Genomics*, **25**, 9–18.

Skuse G.R., Cappione A.J. (1997) RNA processing and clinical variability in neurofibromatosis type I (NF1). *Hum. Mol. Genet*., **6**, 1707–1712.

Thompson J.D., Gibson T.J., Plewniak F., Jeanmougin F., Higgins D.G. (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucl. Acids Res*., **15**, 4876–4882.

Wallace M.R., Marchuk D.A., Andersen L.B., Letcher R., Odeh H.M., Saulino A.M., Fountain J.W., Brereton A., Nicholson J., Mitchell A.L. *et al*. (1990) Type 1 neurofibromatosis gene: identification of a large transcript disrupted in three NF1 patients. *Science*, **249**, 181–186.

MOLECULAR PHYLOGENY OF THE GENUS *TRITICUM* L.

Golovnina K.[1], Glushkov S.[*1], Blinov A.[1], Mayorov V.[2], Adkison L.[2] , Goncharov N.[1]
[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia; [2] Mercer University School of Medicine, Macon, USA
[*] Corresponding author: e-mail: sg23@mail.ru

Key words: wheat, Aegilops, molecular evolution, plasmon and B genome inheritance

SUMMARY

Motivation: The genus *Triticum* L. includes the major cereal crop, common or bread wheat (hexaploid *Triticum aestivum* L.), and other important cultivated species. Wheat has emerged as a classic polyploid model and a significant role of polyploidy as a widespread evolutionary strategy in angiosperms is known. Research on wheat phylogeny has contributed to the understanding of this important phenomenon, but there are still discrepancies and deficiencies in information.

Results: Here, we conducted a phylogenetic analysis of all known wheat species and the closely related Aegilops species. This analysis was based on chloroplast *matK* gene comparison along with *trnL* intron sequences of some species. Polyploid wheat species are successfully divided only into two groups – Emmer (Dicoccoides and Triticum sections) and Timopheevii (Timopheevii section). Results reveal a strictly maternal plastid inheritance of all synthetic wheat amphyploids included in the study. Moreover, a concordance of chloroplast origin with the definite nuclear genomes of polyploid species that were inherited at the last hybridization events was found. This fact allows the most probable donor of the certain nuclear genome and plasmon at the same time to be determined. This suggests that there were two ancestor representatives of *Aegilops speltoides* that participated in the speciation of polyploid wheats with B and G genome in their genome composition. However, G genome species are younger in evolution than ones with B genome and more close to contemporary *Ae. speltoides*.

INTRODUCTION

Species of the genus *Triticum* L. exist as a polyploid series of di-, tetra- and hexaploid wheat with a basic number $n = 7$ and many of these species represent the world's most important food crops. Four basic genomes designated as A, B, D and G assist in the genome constitution of all *Triticum* species (Lilienfeld, Kihara, 1934). Several types of analysis have provided an insight into the ancestry of the definite genomes in allopolyploid species (Zhang *et al*., 2002, Gu *et al*., 2004) and now it is generally accepted that the donors of the A and D genomes were diploid wheats and goatgrass *Aegilops squarrosa* L. (= syn. *Ae. tauschii*), respectively (McFadden, Sears, 1946). However the identity of the donors of the B and G genomes remains open. Many different species have been proposed as the original donor of these genomes but it is now largely believed that the progenitor was a member of the Sitopsis section of the genus Aegilops, namely *Ae. bicornis*, *Ae. longissima*, *Ae. searsii* or, most likely, *Ae. speltoides* (Provan *et al*., 2004). At the same time, there is a hypothesis that the B-genome of polyploid wheat is of a polyphyletic origin, i.e. it is a recombined genome derived from two or more diploid

Aegilops species (Liu *et al*., 2003). In order to elucidate this problem, the results of a molecular phylogenetic analysis including all Triticum species currently described and putative donors of polyploid wheat genomes of the genus Aegilops based on sequence comparison of chloroplast DNA are presented in this study.

METHODS AND ALGORITHMS

Plant materials. Accessions of different wheat species and subspecies, goat grass and rye *Secale cereale* L. and intergeneric synthetics were obtained from N.I. Vavilov All-Russian Institute of Plant Industry (St-Petersburg, Russia), Plant Germ Plasm Institute of Kyoto University (Kyoto, Japan), the National Small Grains Collection (Aberdeen, USA), International Centre for Agriculture Research in the Dry Areas (Aleppo, Syria) and Wageningen Agricultural University (Wageningen, the Netherlands). These species and subspecies include all known wheat species biodiversity. The botanical names of wheat species and their genomic formulas are given according to Goncharov (2005).

Total DNA isolation. Total DNA was isolated as described previously (Rogers, Bendich, 1985).

PCR amplification. The chloroplast *trnL* intron was amplified using a c and d pair of primers (Taberlet *et al*., 1991). A pair of primers for the amplification of the fragment of chloroplast *matK* gene via the polymerase chain reaction (PCR) was designed using sequence of interest of *Triticum aestivum* (GenBank accession number AB042240) Pr2S (5′-CACTTCTCTTTCAGGAATAT-3′) and Pr2A (5′-CATAAAATCGAAGCAAGAGT-3′). All PCR reactions were performed as described previously (Glushkov *et al*., 2006). The PCR products were analyzed in agarose electrophoresis and extracted from gel with a Qiaquick Gel Extraction Kit (Qiagene; according to manufacturer protocol).

DNA sequencing. 200 ng of the PCR product was used in a 10 μl cycle sequencing reaction with the ABI BigDye Terminator Kit on an ABI 377 DNA sequencer. The obtained sequences were deposited to GenBank under accession numbers DQ419998 – DQ420002 (for *trnL* intron), DQ420011 – DQ420046, DQ420048 – DQ420055, DQ436342 (for *matK* gene).

Phylogenetic analysis. The nucleotide sequences were aligned using ClustalX software (Thompson *et al*., 1997). Phylogenetic tree was generated by Neighbor Joining method using MEGA2 software package (Kumar *et al*., 2001). Statistical support for the tree was evaluated by bootstrapping (1000 replications) (Felsenstein, 1985).

RESULTS AND DISCUSSION

Our investigation of the phylogenetic relationships within the genus Triticum consisted of two parts: (i) analysis of the chloroplast *trnL* intron, and (ii) phylogenetic reconstruction based on sequences of *matK* gene.

Analysis of the chloroplast trnL intron. We amplified and sequenced a part of *trnL* intron from some Triticum species belonging to different sections Monococcon (*T. urartu*, *T. monococcum*, *T. boeoticum*), Timopheevii (*T. araraticum*), and Triticum (*T. dicoccoides*). These sequences were analyzed along with obtained from GenBank homologous sequences of *T. aestivum* (AB042240) and other closely related *Ae. tauschii* (AF519113), *Ae. speltoides* (AF519112), and *Ae. uniaristata* (AF519114). There were very few variable sites in the alignment of given sequences. However there was a principal 10 bp insertion (AAACTCATAA) in *Ae. speltoides*, *T. aestivum*, *T. araraticum* and *T. dicoccoides* in the alignment for *trnL* intron that divides all studied species into two groups. We have also distinguished sequences of the diploid Triticum species from the others based on the three specific nucleotide substitutions, which obviously occurred after separation from Aegilops species. This current sequence comparison analysis of *trnL*

intron confirms the hypothesis assuming that the donor of both genomes was an ancestral form of *Ae. speltoides*. Apparently B and G genome of wheats are the offsprings of the *Ae. speltoides* genome with a higher probability than the other diploid Aegilops genomes.

Phylogenetic reconstruction based on sequences of matK gene. We amplified and sequenced the part of *matK* gene for 45 Triticum, Aegilops and other closely related species. Thirty one of 523 aligned sites were variable and an indel characters were coded. We conducted a phylogenetic analysis based on the obtained alignments using the neighbor-joining method (Saitou, Nei, 1987) combined with the molecular evolutionary genetics analysis program MEGA2 (Fig. 1).

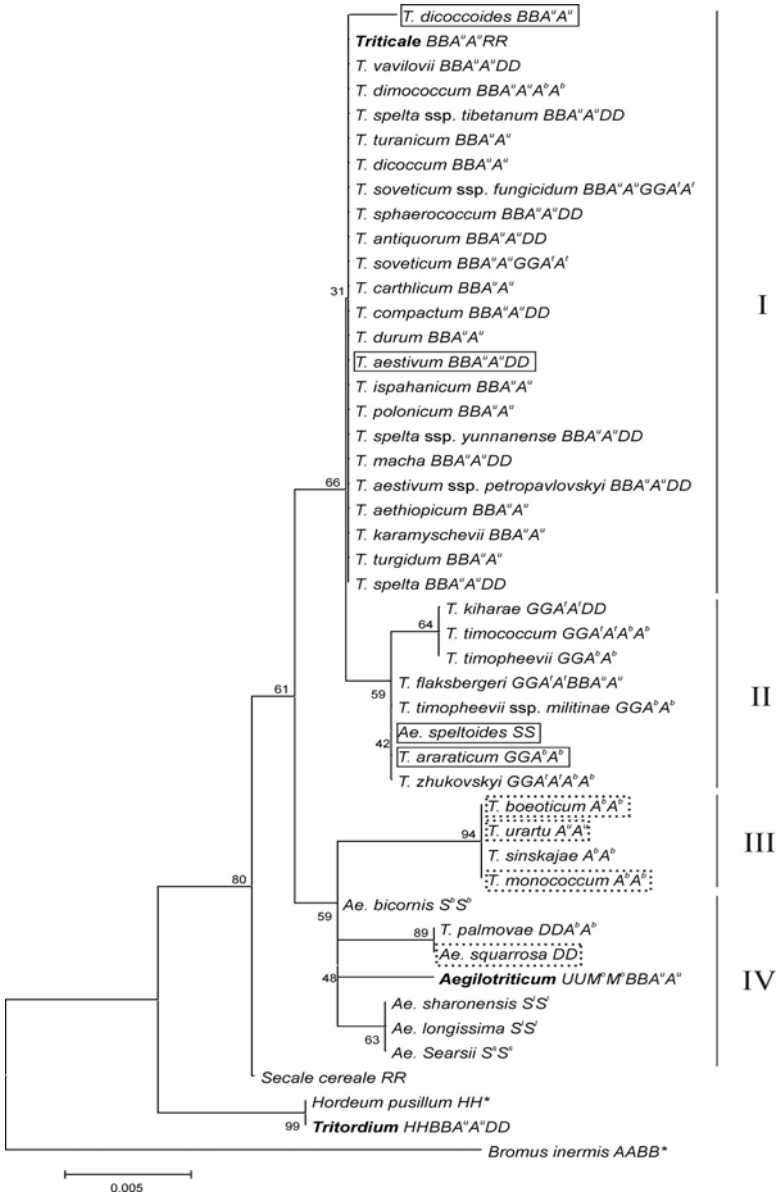


Figure 1. Neighbor-Joining phylogenetic tree based on the comparison of *matK* sequences. Four observed clades are shown by solid lines on the right. There is a genome composition near each species. Synthetic wheats are represented in bold letters. Based on the indel event in the *trnL* intron sequence of some analyzed species, representatives with observed insertions are marked by solid boxes and the rest ones by dotted boxes on the phylogenetic tree. Asteriscs denote species from which the *matK* sequence was obtained from GenBank (AB078133, AF164398).

Sequence of the *matK* gene from *Bromus inermis* was used as an outgroup. Based on the tree topology, all analyzed species except *Secale cereale* and × *Tritordium* are subdivided into four clades of closely related species. The topology of our tree was also supported by a presence of unique substitutions and insertions identified in the *trnL* intron sequences of some Triticum and Aegilops species. Polyploid wheat species are divided into two groups according to the presence of B and G genomes (Fig. 1). As a result, it is concluded that this division occurred after the ancestor of B and G genomes derived from the forefather of A, D and the other studied Aegilops genomes, namely U, S^b, S^l and S^s. Furthermore, *Ae. speltoides* genome S appears most closely related to the B and G genomes.

ACKNOWLEDGEMENTS

The research was financed on the Subprogram II of Program basic research N25 of the Russian Academy of Sciences "The Origin and Evolution of Biosphere" No. 10104-34/P-18/155-270/1105-06-001/28/2006-1.

REFERENCES

Felsenstein J. (1985) Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, **39**, 783–791.

Glushkov S., Novikova O., Blinov A., Fet V. (2006) Divergent non-LTR retrotransposon lineages from the genomes of scorpions (Arachnida: Scorpiones). *Mol. Genet. Genomics*, **275**, 288–296.

Goncharov N.P. (2005) Comparative-genetic analysis – a base for wheat taxonomy revision. *Czech J. Genet and Plant Breed*, **41**, Special issue, 52–54.

Gu Y.Q., Coleman-Derr D., Kong X., Anderson O.D. (2004) Rapid genome evolution revealed by comparative sequence analysis of orthologous regions from four *Triticeae* genomes. *Plant Physiology*, **135**, 459–470.

Kumar S., Tamura K., Jakobsen I.B., Nei M. (2001) MEGA2: molecular evolutionary genetics analysis software. *Bioinformatics*, **17**, 1244–1245.

Lilienfeld F., Kihara H. (1934**)** Genomanalyse bei *Triticum* und *Aegilops*. V. *Triticum timopheevi* Zhuk. *Cytologia*, **6**, 87–122.

Liu B., Segal G., Rong J.K., Feldman M. (2003) A chromosome-specific sequence common to the B genome of polyploid wheat and *Aegilops searsii. Plant. Syst. Evol*., **241**, 55–66.

McFadden E.S., Sears E.R. (1946) The origin of *Triticum spelta* and its free-threshing hexaploid relatives. *J. Hered*., **37**, 81–89.

Provan J., Wolters P., Caldwell K.H., Powell W. (2004) High-resolution organellar genome analysis of *Triticum* and *Aegilops* sheds new light on cytoplasm evolution in wheat. *Theor. Appl. Genet*., **108**, 1182–1190.

Rogers S.O., Bendich A.J. (1985) Extraction of DNA from milligram amounts of fresh, herbarium and mummified plant tissues. *Plant Mol. Biol.*, **5**, 69–76.

Saitou N., Nei M. (1987) The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol*., **4,** 406–425.

Taberlet P., Gielly L., Pautou G., Bouvet J. (1991) Universal primers for amplification of three non-coding regions of chloroplast DNA. *Plant Mol. Biol.*, **17**, 1105–1109.

Thompson J.D., Gibson T.J., Plewniak F., Jeanmougin F., Higgins D.G. (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucl. Acids Res*., **15**, 4876–4882.

Zhang W., Qu L.-J., Gu H., Gao W., Liu M., Chen J., Chen Z. (2002) Studies on the origin and evolution of the tetraploid wheats based on the internal transcribed spacer (ITS) sequences of nuclear ribosomal DNA. *Theor. Appl. Genet*., **104**, 1099–1106.

INFERRING REGULATIORY SIGNAL PROFILES AND EVOLUTIONARY EVENTS

Gorbunov K.Yu.**, *Lyubetsky V.A.
Institute for Information Transmission Problems, RAS, Moscow, 127994, Russia
* Corresponding author: e-mail: gorbunov@iitp.ru

Key words: regulatory signal reconstruction, evolutionary scenario, species tree, frequency matrix, NrdR repression signal, MntR repression signal

SUMMARY

Motivation: Inferring evolution of regulatory signals on the species tree is a timely task. The signal is known at the tips of the tree and is to be reconstructed at its internal nodes.

Results: An algorithm is proposed to reconstruct frequency of every nucleotide and infer evolutionary important edges in the given species tree. Its performance is tested on artificial and biological data including NrdR and MntR regulatory signals. Evolutionary scenarios are inferred for these signals.

THE TASK

Let the species tree be given with its tip taxa assigned multiple alignments of regulatory signals homogeneous across species in each taxon. NrdR is one of such signals (for an example of such types to Rodionov, Gelfand, 2005). Each alignment of n columns produces a corresponding $4 \times n$ frequency matrix of 4 nucleotide letters and n columns. The task is to reconstruct corresponding frequency matrices at internal nodes, as well as, for each column i of the signal, to infer edges in the species tree containing important events in the signal evolution. Note that i also designates i-columns of frequency matrices generated at the tips. The edge is considered *evolutionary important* if maximum parsimony requirement for it is violated in the sense that frequency matrices reconstructed at its nodes (and their i-columns) differ considerably. Evolutionary patterns at i- and j-positions of the signal can be different. Therefore, evolutionary scenarios are reconstructed for each position separately and then juxtaposed. Under fixed i-position in each leaf, i-column in the corresponding frequency matrix is defined (it is called the signal profile or frequency *distribution* at i-position). The aim is, to reconstruct such distributions in each inner node under fixed i-position and infer edges containing evolutionary important events for i-position.

Such scenarios for a position are defined as *individual scenarios* and are further joined in a *resulting scenario*, which combines the edges contributing the most into evolution of the entire signal. Also, *substantial positions* are inferred as those having relatively robust evolutionary scenarios with respect to the signal structure.

ALGORITHM

Maximum parsimony is used to solve this task. Namely we minimize function F, which is the sum of pairwise distances between distributions at adjacent nodes. Two

conditions are imposed: the sum of fractions at each node gives 1, all fractions are positive. Although our algorithm allows for other functions F and other distribution conditions as well. For edge u the corresponding sum of four items in function F is denoted $F(u)$. An iterative step is as follows: F is minimized, and edges with highest values $F(u)$ are considered in the number determined by parameter *vet*. For *each such u*, item $F(u)$ is independently subtracted from F, thus hypothesizing edge *u to contain an evolutionary event* and therefore not be maximally parsimonious, after which resulting F is again minimized.

The procedure iterates until the number of excluded edges exceeds the value, determined by the *glub* parameter. Each succession of excluded edges is called the evolutionary *i*-scenario under given *vet* and *glub* (*branching* and *depth*, respectively). The settings in test run were *vet* = 15 and *glub* = 4. A *robust scenario* is defined with a combination of three requirements: lower number of its contained edges, lower $F(u)$ value for all non-excluded edges, lesser sum of pairwise distances between distributions at non-excluded edges and higher – at excluded ones. The algorithm also implements comparative analysis of different *i*-scenarios to reveal their consistency (i.e. presence of an evolutionary event at the same edge in several corresponding positions of the signal) and robustness in individual signal positions (those are called evolutionary important positions). Evolutionary not important positions are excluded and the algorithm is applied to the rest of signal. The *resulting evolutionary scenario* includes edges from different individual scenarios, especially those shared by *i*-scenarios of coevolving positions (e.g., direct or inverted repeats) and which are robust for many positions. The edge contributes in the resulting scenario if contained in several robust *i*-scenarios, with weight of corresponding *i*-positions being high.

RESULTS AND DISCUSSION

Here we describe the algorithm's performance on two biological datasets. NrdR-box of length 16 (Rodionov and Gelfand, 2005) is involved in biosynthesis regulation of replication-associated molecules. The species tree used in the study is shown in Fig. 1 (edges designated with numbers of their corresponding descendant nodes). For 16 signal positions our algorithm found the following robust scenarios: (1) 40, 6, 12, 16; (2) 40, 30, 26, 23; (3) 2, 17, 29, 25; (4) 2, 4, 11, 13; (5) 40, 2, 12, 16; (6) 2, 3, 6, 16; (7) 30, 26, 7, 37; (8) 40, 2, 4, 3; (9) 40, 2, 3, 15; (10) 40, 30, 26, 16; (11) 40, 2, 17, 18; (12) 26, 31, 39, 16; (13) 40, 2, 3, 13; (14) 40, 2, 3, 17; (15) 40, 2, 12, 16; (16) 40, 2, 5, 32. The resulting scenario is combined from edges 40 and 2, which suggests considerable changes in NrdR signal to have happened during its evolution in this part of the tree.

The recently discovered MntR-box serves as the second example (Mn transport regulation, Rodionov D.A., Gelfand M.S., 2006, personal communication). The species tree is given in Fig. 2. For 22 signal positions the algorithm output 22 robust scenarios: (1) 20, 24, 18, 14; (2) 1, 6, 11, 12; (3) 1, 6, 7, 9; (4) 2, 4, 24, 21; (5) 2, 24, 6, 21; (6) 1, 3, 6, 10; (7) 1, 4, 6, 10; (8) 1, 2, 4, 14; (9) 1, 2, 4, 24; (10) 1, 24, 10, 13; (11) 2, 24, 21, 18; (12) 6, 9, 22, 13; (13) 1, 21, 10, 11; (14) 4, 24, 21, 14; (15) 1, 4, 6, 14; (16) 1, 4, 6, 10; (17) 1, 6, 7, 10; (18) 1, 2, 24, 21; (19) 1, 2, 4, 13; (20) 1, 6, 17, 14; (21) 1, 3, 9, 11; (22) 2, 4, 21, 22. The resulting scenario is combined from edges 1 and 6, thus suggesting phylogenetic localization of the signal change over time.

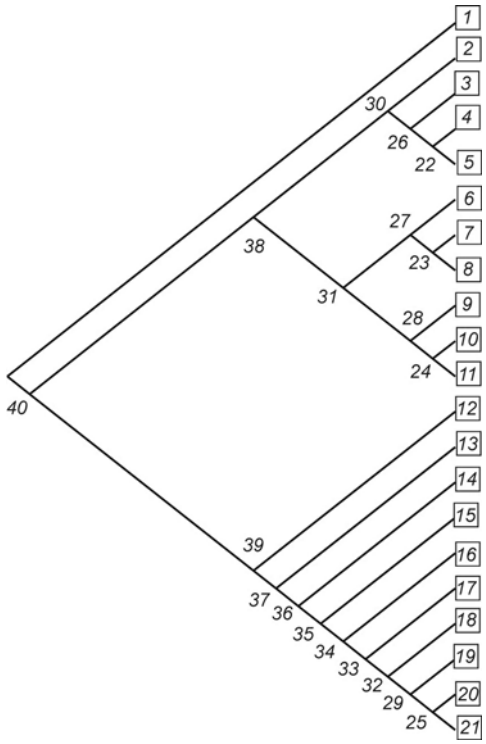


Figure 1. Species tree for the case of NrdR-box with consensus acaC(a/t)AtATaT(a/t)Gtg.

Taxa designations in Fig. 1 are as follows: 1 = {*T. maritima*, *T. thermophilus*}; 2 = {*D. radiodurans*}; 3 = {*P. marinus*, *G. violaceus*, *Synechocystis sp.*, *S.elongates*, *T. elongates*}; 4 = {*S. coelicolor*, *S. avermitilis*, *S. scabies*, *C. michiganensis, L. xyli*, *Corynebacterium* spp., *Mycobacterium* spp.}; 5 = {*P. acnes*, *B. longum*, *T. fusca*}; 6 = {*S. aureus*}; 7 = {*C. acetobutylicum*, *C. tetani*, *C. perfringens*, *C. botulinum*, *C. difficile*, *T. tengcongensis*, *C. hydrogenoformans*, *D. hafniense*}; 8 = {*B. subtilis*, *B. licheniformis*, *B. halodurans*, *B. cereus*, *B. stearothermophilus*}; 9 = {*E. faecalis*, *E. faecium*}; 10 = {*S. epidermidis*, *S. pyogenes*, *S. agalactiae*, *S. pneumoniae*, *S. mutans*, *P. pentosaceus*}; 11 = {*Lactobacillus* spp.}; 12 = {*C. muridarum*, *C. pneumoniae*, *C. trachomatis*, *C. abortus*, *C. caviae*, *T. denticola*}; 13 = {*G. sulfurreducens*, *G. metallireducens*, *D. acetoxidans*, *D. psychrophila*, *B. bacteriovorans*, *B. marinus*, *M. xanthus*}; 14 = {*B. melitensis*, *M. loti*, *A. tumefaciens*, *R. leguminosarum*, *S. meliloti*, *B. japonicum*, *R. palustris*, *R. capsulatus*, *C. crescentus*, *H. neptunium*, *E. chaffeensis*, *N. sennetsu*}; 15 = {*N. europaea*, *N. meningitides*, *M. flagelatus*, *R. solanacearum*, *B. pertussis*, *B. bronchseptica*, *B. avium*, *B. fungorum*, *B. cepacia*, *B. pseudomallei*, *D. aromatica*}; 16 = {*X. fastidiosa*, *X. axonopodis*}; 17 = {*P. aeruginosa*, *P. putida*, *P. fluorescens*, *P. syringae*}; 18 = {*V. cholerae*, *V. vulnificus*, *V. parahaemolyticus*}; 19 = {*E. coli*, *S. typhi*, *K. pneumoniae*, *Y. pestis*, *Y. enterocolitica*, *E. chrysanthemi*, *E. carotovora*, *P. luminescens*}; 20 = {*P. multocida*}; 21 = {*H. influenzae*, *H. ducreyi*}.

Taxa designations in Fig. 2 are as follows: 1 = {*T. fusca*, *R. xylanophilus*, *C. diphtheria*, *C. efficiens*, *C. glutamicum*}; 2 = {*Streptococcus* spp., *L. lactis*, *P. filamentus*, *C. hutchinsonii*}; 3 = {*E. faecalis*}; 4 = {*B. subtilis*, *B. cereus*, *B. halodurans*, *O. iheyensis*, *L. monocytogenes*}; 5 = {*Staphylococcus* spp.}; 6 = {*Treponema* spp.}; 7 = {*M. magnetotacticum*, *R. capsulatus*, *Mesorhizobium* spp.}; 8 = {*E. coli*, *S. typhi*, *K. pneumoniae*}; 9 = {*Xanthomonas* spp., *X. fastodiosa*}; 10 = {*Methanosarcina* spp.}; 11 = {*A. fulgidus*}; 12 = {*M. thermophila*}; 13 = {*Pyrococcus* spp.}; 14 = {*M. jannaschii*, *M. maripaludis*}.

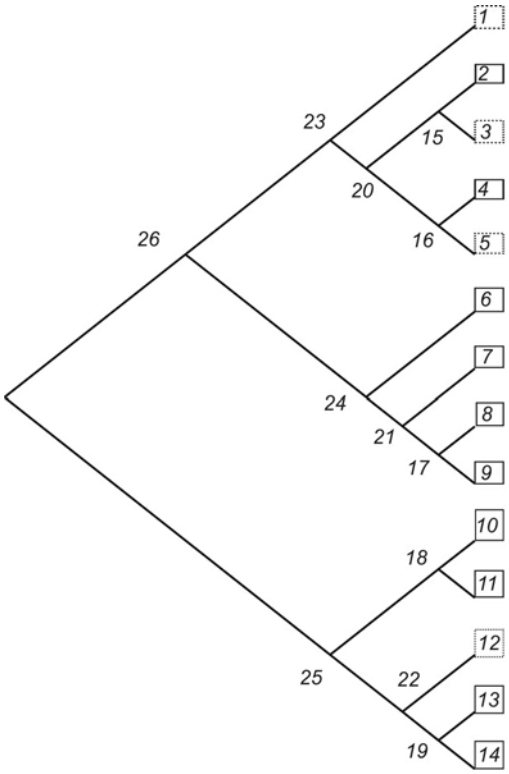


Figure 2. Species tree for the case of MntR-box with consensus a(a/t)(a/t)TTTAG(c/g)nnnn(g/c) ctA Aa(a/t)(a/t)n.

ACKNOWLEDGEMENTS

The authors are greatly indebted to prof. M.S. Gelfand for fruitful discussions, D.A. Radionov for providing test data and discussions, to L.Y. Rusin and A.V. Seliverstov for valuable help and discussions.

REFERENCES

Rodionov D.A., Gelfand M.S. (2005) A universal regulatory system of ribonucleotide reductase genes in bacterial genomes. *Trends in Genet.*, **21,** 385–398.

A METHOD FOR SEMIAUTOMATED ANALYSIS OF GENE EVOLUTION

Gunbin K.V.*[*], *Morozov A.V., Afonnikov D.A.
Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia
[*] Corresponding author: e-mail: genkvg@bionet.nsc.ru

Key words: multiple alignment, phylogenetic trees, adaptive evolution

SUMMARY

Motivation: With increasing the amount of data on primary structures of genetic macromolecules it is very important to develop new methods for the large-scale evolutionary analysis. There exist a large number of programs aimed at solving of different types of evolutionary analysis, which are freely distributed. However, the computational tools often solve a very specific task. This makes difficult to implement the full cycle of analysis for the large scale datasets. Thus, the developing multitask and integrative approaches is very important.

Results: The multitask system for solving a wide range of evolutionary analysis problems is developed by integration of a set of programs for evolutionary analysis. All the programs are united to an adaptable pipeline and allow solving different task of analysis from multiple alignment to analysis of adaptive evolution of the genes.

Availability: The multitask system is written in Perl and is available from the authors upon request.

INTRODUCTION

Current sequencing projects allow accumulating the huge amount of data on DNA, RNA and protein primary structures at a high rate. As a result, new information is available from comparative analysis of such a data (Bateman *et al.*, 2004; Roth *et al.*, 2005; Whelan *et al.*, 2006). The obtaining such an information is usually a multitask problem that utilize several step of widely admitted problems such as search for homologous sequences, multiple sequence alignment, phylogenetic tree reconstruction, detection and analysis of adaptive evolution and so forth. There are many tools aimed at solving each particular task using different models and approximations and producing results in different data formats that should be combined by biologist. Therefore, the technical task of integration of results at different steps of analysis may need a lot of time. From the other hand, it is can also be important to researcher to compare results from different approaches to make its combination that can be more biologically relevant. It is important to note, that such an integrative analysis would be more useful if implemented in automatic manner allowing processing large sets of protein families that are available from databases (Bateman *et al.*, 2004).

In this work we suggest semi-automated multitask program package that can implement and integrate the variety of free-available programs. The advantage of our system is that it allows to make user-defined pipelines of evolutionary analysis of biological sequences, select and combine the obtained results at the specific analysis step.

METHODS AND ALGORITHMS

The pipeline of the semiautomated analysis of gene evolution ways consists of five steps: 1) a search for the amino acid and nucleotide sequences, 2) multiple sequence alignment; 3) constructing phylogenetic trees; 4) a search for adaptively evolving gene regions and 5) a search for the phylogenetic tree branches where adaptive evolution has presumably been the case. The implementation of each step is described below.

1. User-controlled searches of the GenBank (Pruitt *et al.*, 2005) and Ensembl (Hubbard *et al.*, 2005) databases for amino acid sequences and their encoding nucleotide sequences is performed using PSI-BLAST. The sequences retrieved automatically, then user-assisted selection is performed on the basis of formal properties: sequence length; integrity; degree of homology with a queried sequence.

2. Multiple alignment of the amino acid sequences is performed in several steps: i) multiple alignments are made using the CLUSTALW (Thompson *et al.*, 1994), T-COFFEE (Notredame *et al.*, 2000), MUSCLE (Edgar, 2004) and MAFFT (Katoh *et al.*, 2005) software programs; ii) multiple alignments are refined with an automatic procedure performed using the LEON (Thompson *et al.*, 2004) and RASCAL (Thompson *et al.*, 2003) program packages; iii) using a set of PERL programs and the EMBOSS program package (Rice *et al.*, 2000), the weights of the multiple alignments of protein sequences is calculated and they are compared on the basis of information in the ProSite (Hulo *et al.*, 2004) and NCBI CDD databases (Marchler-Bauer *et al.*, 2005). If deemed necessary, the multiple alignment can be refined manually. Automatic DNA sequence alignment is performed on the basis of the protein sequence alignments.

3. Before proceeding to phylogeny reconstruction, a search is performed for an optimal model for protein or DNA evolution using the PROTTEST (Abascal *et al.*, 2005) and MODELTEST (Posada, Crandall, 1998) programs. The model, which is optimal for protein and/or DNA evolution is used for phylogeny reconstruction using the PHYML (Guindon, Gascuel, 2003) and TREE-PUZZLE (Schmidt *et al.*, 2002). A user-assisted check is performed for the reconstructed phylogenetic trees scale up to realistic trees based on data in the TreeBase (Morell, 1996), NCBI Taxonomy (Wheeler *et al.*, 2000) and Tree of Life (Maddison, Schulz, 2005) databases. If deemed necessary, the phylogenetic trees can be refined manually.

4. The search for adaptively evolving gene regions can be done using a variety of methods: by the method used in the PSWIN program (Whittam, Pennsylvania State University) or by the CRANN program (Creevey, McInerney, 2003); by nucleotide sequences using maximum likelihood without simulating evolution (PAML (Yang, 1997)) and using Monte Carlo simulation of evolution (PLATO (Grassly, Holmes, 1997)) or by amino acid sequences, using RATE4SITE (Pupko *et al.*, 2002) program. The profiles of the adaptive evolution developed using the different techniques are superimposed on one another along the coding sequence.

5. The search for the adaptive evolution mode along the tree branches uses pair-wise comparison of sequences or analysis of the way of gene evolution by reconstruction of ancestral gene sequences. Pair-wise comparison is performed using the NEW1 and NEW2 (Ina, 1995), KHKS (Tang, Wu, 2006), CRANN (Creevey, McInerney, 2003) software programs and the PAML program package (Yang, 1997). Both absolute values of K_a/K_s (ratios of synonymous to non-synonymous nucleotide substitutions) and the ratios between them are taken into account. Analysis of the way of gene evolution by reconstruction of ancestral gene sequences is performed using the CRANN (Creevey, McInerney, 2003) software program and the Norwegian Bioinformatics Platform server (Siltberg, Liberles, 2002). The topology-specific patterns of the adaptive evolution of the genes (by looking at the phylogenetic tree) identified in user-assisted mode.

RESULTS

The evolution of the genes in the Hedgehog, Decapentaplegic, Wingless and Notch signaling cascades have been analyzed using the method described herein. It has been revealed that there is a correlation between the adaptive evolution of transcription factors, morphogens, their ontogenesis-related receptors and aromorphoses (Gunbin *et al.*, 2005a, b; 2006).

ACKNOWLEDGEMENTS

The Project *The evolution of molecular-genetic systems: computer analysis and modeling* of the Program *Biosphere Origin and Evolution* of the Presidium of the Russian Academy of Sciences No. 10104-34/P-18/155-270/1105-06-001/28/2006-1, RFBR No. 05-07-98012 *Development of a GRID-based computational portal for bioinformatics*, the Innovation project of Federal Agency of Science and innovation IT-CP.5/001 "Development of software for computer modeling and design in postgenomic system biology (system biology *in silico*)".

REFERENCES

Abascal F. *et al.* (2005) ProtTest: selection of best-fit models of protein evolution. *Bioinformatics*, **21**, 2104–2105.

Bateman A. *et al.* (2004) The Pfam protein families database. *Nucl. Acids Res.*, **32**, D138–D141.

Creevey C.J., McInerney J.O. (2003) CRANN: detecting adaptive evolution in protein-coding DNA sequences. *Bioinformatics*, **19**, 1726.

Edgar R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 113.

Grassly N.C., Holmes E.C. (1997) A likelihood method for the detection of selection and recombination using nucleotide sequences. *Mol. Biol. Evol.*, **14**, 239–247.

Guindon S., Gascuel O. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.*, **52**, 696–704.

Gunbin K.V. *et al.* (2005a) Aromorphoses and the adaptive molecular evolution: hedgehog signaling cascade genes. *International Workshop Biosphere Origin and Evolution,* Novosibirsk, Russia, June 26-29, 2005.

Gunbin K.V. *et al.* (2005b) Aromorphoses and the adaptive molecular evolution: hedgehog and wingless aignaling cascades genes. *The 2005 BGRS International Summer School for young scientists "Evolution, Systems Biology and High Performance Computing Bioinformatics",* Novosibirsk, Russia, September 11-16, 2005.

Gunbin K.V. (2006) Aromorphoses and adaptive molecular evolution: morphogens and signaling cascade genes. *This issue*.

Hubbard T. *et al.* (2005) Ensembl 2005. *Nucl. Acids Res.*, **33**, D447–D453.

Hulo N. *et al.* (2004) Recent improvements to the PROSITE database. *Nucl. Acids Res.*, **32**, D134–D137.

Ina Y. (1995) New methods for estimating the numbers of synonymous and nonsynonymous substitutions. *J. Mol. Evol.*, **40**, 190–226.

Katoh K. *et al.* (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucl. Acids Res.*, **33**, 511–518.

Maddison D.R., Schulz K.-S. (ed.) 2005. The Tree of Life Web Project. Internet address: http://tolweb.org

Marchler-Bauer A. *et al.* (2005) CDD: a conserved domain database for protein classification. *Nucl. Acids Res.*, **33**, D192–D196.

Morell V. (1996) TreeBASE: the roots of phylogeny. *Science*, **273**, 569.

Notredame C. *et al.* (2000) T-COFFEE: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, **302**, 205–217.

Posada D., Crandall K.A. (1998) MODELTEST: testing the model of DNA substitution. *Bioinformatics*, **14**, 817–818.

Pruitt K.D. *et al.* (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucl. Acids Res.*, **33**, D501–D504.

Pupko T. *et al.* (2002) Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics*, **18**, S71–S77.

Rice P. *et al.* (2000) EMBOSS: the european molecular biology open software suite. *Trends Genet.*, **16**, 276–277.

Roth C. *et al.* (2005) The adaptive evolution database (TAED): a phylogeny based tool for comparative genomics. *Nucl. Acids Res.*, **33**, D495–D497.

Schmidt H.A. *et al.* (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics*, **18**, 502–504.

Siltberg J., Liberles D.A. (2002) A simple covarion-based approach to analyse nucleotide substitution rates. *J. Evol. Biol.*, **15**, 588–594.

Tang H., Wu C.-I (2006) A new method for estimating nonsynonymous substitutions and its applications to detecting positive selection. *Mol. Biol. Evol.*, **23**, 372–379.

Thompson J.D. *et al.* (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl. Acids Res.*, **22**, 4673–4680.

Thompson J.D. *et al.* (2003) RASCAL: rapid scanning and correction of multiple sequence alignments. *Bioinformatics*, **19**, 1155–1161.

Thompson J.D. *et al.* (2004) LEON: multiple alignment evaluation of neighbours. *Nucl. Acids Res.*, **32**, 1298–1307.

Yang Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.*, **13**, 555–556.

Wheeler D.L. *et al.* (2000) Database resources of the national center for biotechnology information. *Nucl. Acids Res.*, **28**, 10–14.

Whelan S. *et al.* (2006) PANDIT: an evolution-centric database of protein and associated nucleotide domains with inferred trees. *Nucl. Acids Res.*, **34**, D327–D331.

AROMORPHOSES AND ADAPTIVE MOLECULAR EVOLUTION: MORPHOGENS AND SIGNALING CASCADE GENES

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

Key words: morphogens, transcription factors, receptors, protein families, adaptive evolution

SUMMARY

Motivation: Molecular evolution data can help the researchers understand the evolutionary process when there is no palaeontological information about taxa in question. The molecular evolution of signal cascade components is particularly interesting when we try to analyse the early evolution of Metazoa because: 1) the fossil records of morphologically intermediate taxa, transitional forms between contemporary metazoan taxa and Ediacara fauna, are scanty (Peterson *et al.*, 2000) and 2) the particular signal cascades are present only in Metazoan taxa (Pires-daSilva, Sommer, 2003).

Results: The evolutionary modes of proteins morphogens, their receptors and key transcription factors involved in the Wingless, Hedgehog and Decapentaplegic signaling cascades have been analyzed. It has been revealed that there is a correlation between the timing of the adaptive evolution of signal cascade components and timing of Metazoan aromorphoses[1].

INTRODUCTION

In 2000, Duret and Mouchiroud in their work on the potencies of various groups of mammalian genes to fix non-synonymous nucleotide substitutions demonstrated that the higher gene functional load, the less likely the fixation of such substitutions (Duret, Mouchiroud, 2000). We are searching for such events of adaptive evolution that could be attributed to the earlier stages of many-celled animal evolution. For that purpose, we looked into the evolution of genes responsible for the embryonic development of many-celled animals. The analysis covered the genes encoding morphogens, their receptors and key transcription factors involved in the Wingless-, Hedgehog- and Decapentaplegic signaling cascades, which are by far the most important components in animal tissue differentiation processes (Pires-daSilva, Sommer, 2003; Gunbin *et al.*, 2004).

METHODS

A search for homologous sequences, multiple sequence alignments and phylogenetic tree reconstruction were performed as described in Gunbin *et al.*, 2006 (Gunbin *et al.*, 2006). In the search for homologous sequences procedure the entries were only

[1] Adaptive changes in the organisms increasing total fitness.

experimentally annotated proteins of vertebrate and invertebrate species. The search for the gene regions that undergo adaptive evolution and the branches of phylogenetic trees on which adaptive evolution has ever occurred was made done only among the species that were close enough to the roots of the sub-trees being examined.

The search for adaptively evolving gene regions was made using a variety of techniques: 1) applied to nucleotide sequences – using the PSWIN (Whittam, Pennsylvania State University), using PAML (Yang, 1997) and PLATO (Grassly, Holmes, 1997); 2) applied to amino acid sequences using RATE4SITE (Pupko *et al.*, 2002). The gene region under adaptive evolution considered reliable if all methods with the exception of one found that region.

The search for phylogenetic tree branches on which gene adaptive evolution has occurred was made among the gene regions that had been identified at the previous stage as being adaptively evolving. Only pair-wise sequence comparisons were carried out. Analysis was performed using the K-ESTIMATOR (Comeron, 1999) (Fig. 1, C95), KHKS (Tang, Wu, 2006) (Fig. 1, TW06_Ka and TW06_Kh) and MEGA3 (Kumar *et al.*, 2004) programs (Fig. 1, NG86m) and the PAML (Yang, 1997) program package (Fig. 1, YN00). The number of synonymous transversions (DAMBE applied (Xia, Xie, 2001)) and pair-wise mismatches in the alignments of amino acid sequences (TREE-PUZZLE applied (Schmidt *et al.*, 2002)) was calculated in order to get the WAG/NED values (Fig. 1). In so doing, only the ratios of $(K_a/K_s)_i$ to $(K_a/K_s)_j$ values were regarded, not the absolute values of K_a/K_s[2] (Fig. 1) (Gunbin *et al.*, 2005b; unpublished data).

RESULTS

The evolution of the following genes was analyzed: 1) the genes *Hh*, *Ptc* and *Ci* of the Hedgehog signaling cascade; 2) the genes *Dpp*, *Tkv*, *Put* and *Mad* of the Decapentaplegic signaling cascade; 3) the genes *Wg*, *Fz* and *Pan* of the Wingless signaling cascade. Over 700 nucleotide and amino acid sequences have been analyzed.

We have demonstrated that the adaptive evolution of *Hh* (Fig. 1), *Dpp* and *Wg* correlates with the emergence of the arthropods and vertebrates and the groups of paralog genes in the vertebrates (Gunbin *et al.*, 2005a, b). We have demonstrated that the adaptive evolution of *Hh*, *Dpp* and *Wg* is confined to the regions that are responsible for the rate at which the morphogen forms and/or the specificity of the morphogen for the receptor.

The adaptive evolution of the genes *Ptc*, *Fz* encoding transmembrane proteins receptors, correlate with the emergence of large taxa of many-celled organisms: the arthropods and the vertebrates (Gunbin *et al.*, 2005a, b). For example, in the protein Ptc, adaptively evolving is the sterol-sensitive domain which is required for interacting with the downstream co-receptor, the protein Smo (Gunbin *et al.*, 2005a, b). The proteins Fz, Tkv and Put, which function as receptors of the morphogens Wg and Dpp, evolve likewise; they do so where they are in contact with the components of the signaling cascade within the cell cytoplasm.

We have clearly demonstrated that the adaptive evolution of genes encoding the transcription factors Ci, Mad and Pan correlates with the emergence of large taxa of bilateral organisms (arthropods and vertebrates). The adaptive evolution of proteins Ci, Mad and Pan focus on the domains where they are in contact with the transcription cofactors (Gunbin *et al.*, 2005a, b). For example for Ci protein it is highly statistically significant that evolution goes on at especially high rates in the domain responsible for binding to transcription cofactors (Gunbin *et al.*, 2005a, b).

[2] K_a/K_s – the ratio of the number of non-synonymous to synonymous nucleotide substitutions.

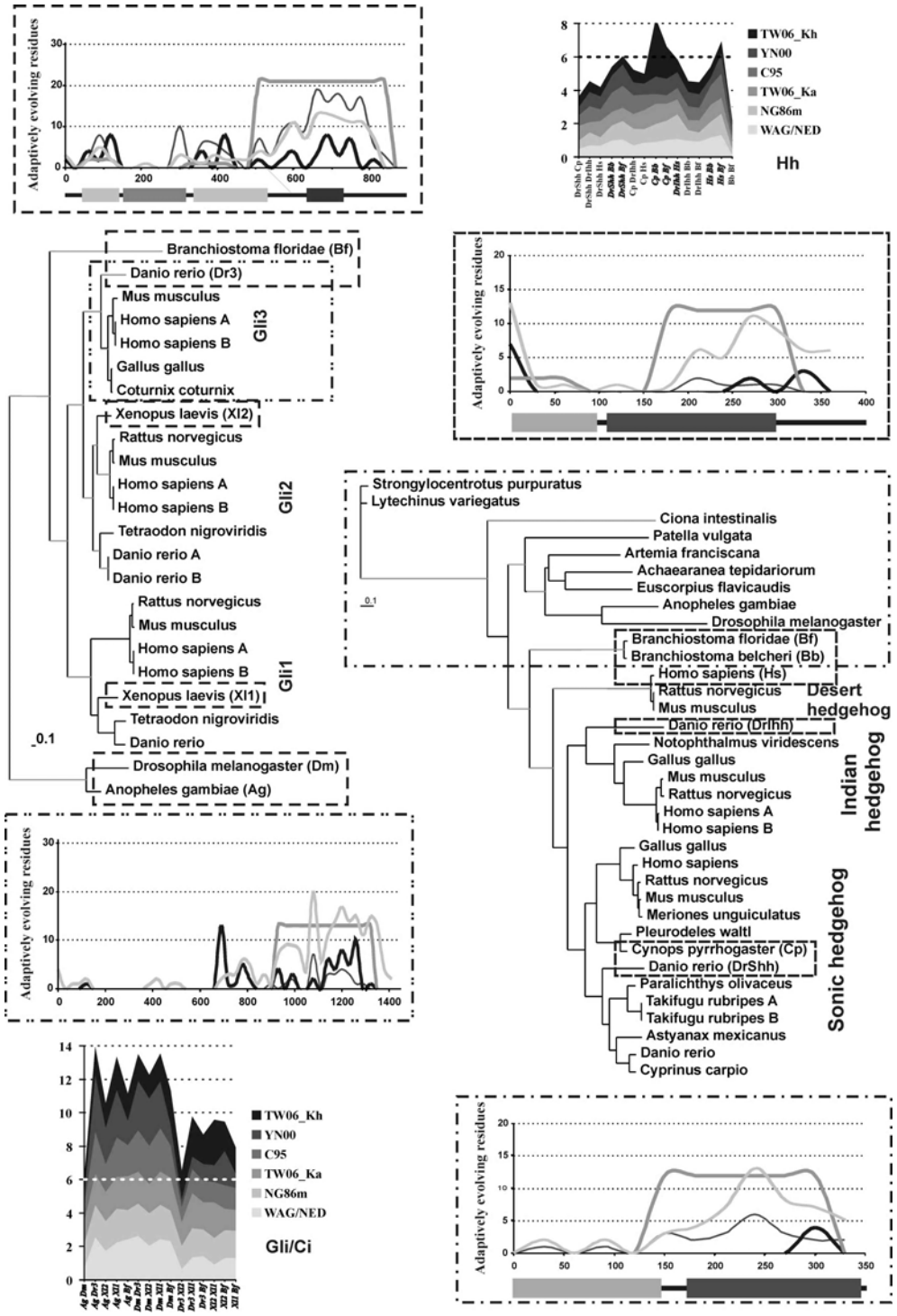


Figure 1. Adaptive evolution of the Ci (left) and Hh (right) family genes. Gray branches on the phylogenetic trees are where adaptive evolution is asserted to be a fact. Specific dotted rectangles on the phylogenetic trees run around the species which have genes in question. The profiles presented on the scatter plots display the evolutionary lability of the proteins encoded by the genes in question. Filled rectangles below the scatter plots indicate conserved protein regions. The profiles presented on the stacked area plots display the relative values of K_a/K_s for the specific pair-wise comparisons of genes in question. On the scatter plots: PLATO (dark-gray heavy line), RATE4SITE (light-gray heavy line), PAML (black heavy line), PSWIN (black light line). On the x axis at the scatter plots: the numeration of amino acids on the multiple alignment of proteins; begins at the N-end; on the y axis: the number of adaptively evolving

amino acid residues per frame (for PAML and PSWIN); for PLATO, Z-score is significantly different from random. On the x axis at the stacked area plots – the sum of relative values of K_a/K_s (6 – critical level); on the y axis – the notations of specific pair-wise comparisons (the notations of the comparisons correspond to two species short names which have genes in question). From the right to the stacked area plots are shorthand notations for the pair-wise sequence comparisons methods applied (see text).

We have therefore demonstrated that the adaptive evolution of transcription factors, morphogens and their ontogenesis-related receptors correlate with aromorphoses. This could be explained assuming that at the earliest stages of evolution of bilateral organisms the genes that are now responsible for entire ontogenesis were only responsible for one of its stages.

ACKNOWLEDGEMENTS

This work was in part supported by the project *Evolution of Molecular Genetic Systems: Computer Analysis and Modeling* of the program *Biosphere Origin and Evolution* of the Presidium of the Russ. Acad. Sci. No. 10104-34/P-18/155-270/1105-06-001/28/2006-1 and by grant No. 05-07-98012 from RFBR *Development of the GRID-based Computational Portal for Bioinformatics*.

REFERENCES

Comeron J.M. (1999) K-Estimator: calculation of the number of nucleotide substitutions per site and the confidence intervals. *Bioinformatics*, **15**, 763–764.

Duret L., Mouchiroud D. (2000) Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. *Mol. Biol. Evol.*, **17**, 68–74.

Grassly N.C., Holmes E.C. (1997) A likelihood method for the detection of selection and recombination using nucleotide sequences. *Mol. Biol. Evol.*, **14**, 239–247.

Gunbin K.V. *et al.* (2004) Genetic mechanisms of morphological evolution, part 1. *Sib. J. of Ecol.*, **11**, 611–621. (In Russ.).

Gunbin K.V. *et al.* (2005a) Aromorphoses and the adaptive molecular evolution: hedgehog signaling cascade genes. Oral presentation. *International Workshop "Biosphere Origin and Evolution",* Novosibirsk, Russia, June 26-29, 2005.

Gunbin K.V. *et al.* (2005b) Aromorphoses and the adaptive molecular evolution: hedgehog and wingless aignaling cascades genes. Oral presentation. *The 2005 BGRS International Summer School for young scientists "Evolution, Systems Biology and High Performance Computing Bioinformatics",* Novosibirsk, Russia, September 11-16, 2005.

Gunbin K.V. *et al.* (2006) A method for semiautomated analysis of gene evolution. *This issue.*

Kumar S. *et al.* (2004) MEGA3: Integrated software for molecular evolutionary genetics analysis and sequence alignment. *Brief Bioinform.*, **5**, 150–163.

Peterson K.J. *et al.* (2000) Bilaterian origins: significance of new experimental observations. *Dev. Biol.*, **219**, 1–17.

Pires-daSilva A., Sommer R.J. (2003) The evolution of signalling pathways in animal development. *Nat. Rev. Genet.*, **4**, 39–49.

Pupko T. *et al.* (2002) Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics*, **18**, S71–S77.

Schmidt H.A. *et al.* (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics*, **18**, 502–504.

Tang H., Wu C.-I (2006) A new method for estimating nonsynonymous substitutions and its applications to detecting positive selection. *Mol. Biol. Evol.*, **23**, 372–379.

Xia X., Xie Z. (2001) DAMBE: software package for data analysis in molecular biology and evolution. *J. Hered.*, **92**, 371–373.

Yang Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.*, **13**, 555–556.

EVOLUTIONARY RELATIONSHIPS AND DISTRIBUTION OF THE DIFFERENT LTR RETROTRANSPOSON FAMILIES IN PLANTS

Kabanova A., Novikova O.[1], Gunbin K.[1], Fet V.[2], Blinov A.[*1]
[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia; [2] Marshall University, Huntington,WV, USA
[*] Corresponding author: e-mail: blinov@bionet.nsc.ru

Key words: retrotransposons, LTRs, plants, evolution

SUMMARY

Motivation: LTR retrotransposons are the common class of transposable elements found in plant genomes. They can comprise a huge fraction of certain genomes and may play important roles in genome organization and maintenance of genome fitness. Phylogenetic analysis and distribution of LTR retrotransposons in the different plant taxa can give us insights not only into the co-evolution of LTR retrotransposons and hosts, but also indicate some perspectives in retrotransposon research.

Results: We investigated a distribution of Ty3/gypsy and Ty1/copia elements in 19 different species from 17 plant taxa. Totally, 169 clones with reverse transcriptase sequences were determined. LTR retrotransposons have been found for the species, in which no retroelements had been earlier described. That includes Bryophyta, Lycopodiophyta, Sciadopytiaceae, Nympheales, Winterales and Piperales in case of Ty1/copia elements; and Bryophyta, Lycopodiophyta, Filicophyta, Cycadophyta, Sciadopytiaceae, Nympheales, Winterales and Piperales in case of Ty3/gypsy elements. Several different families of both Ty3/gypsy and Ty1/copia elements were determined.

INTRODUCTION

Retrotransposons are mobile genetic elements that propagate themselves by reverse transcription of an RNA intermediate. There are two major classes of retrotransposons: LTR retrotransposons present 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 carry terminal repeats and employ a simpler target-primed reverse transcription (TPRT) mechanism for retrotransposition. Retrotransposons are ubiquitous among plants and play sufficient role in the evolution of their genes and genomes because of the broad spectrum of mutations produced by their activity (Kidwell, Lisch, 2000). The two large families, Ty1/copia and Ty3/gypsy, of LTR retrotransposons were described in plants. Ty3/gypsy elements have a structural similarity to retroviruses, whereas Ty1/copia elements essentially differ from them. These results demonstrated the convenience for classifying LTR retroelements as viruses. Ty1/copia elements were referred to Pseudoviridae family and Ty3/gypsy elements, to Metaviridae family.

METHODS AND ALGORITHMS

Total DNA isolation. Total DNA was isolated using a standard CTAB method (Rogers, Bendich, 1985).

PCR amplification and sequencing. Based on the comparison of 72 sequences of the known plant LTR retrotransposon reverse transcriptase domains, degenerate oligonucleotides primers were constructed using CODEHOPE software (Rose *et al.*, 1998) to amplify a 400–450 bp region for Ty1/copia elements, 300–350 bp for Ty3/gypsy elements, and nearly 350 bp for Athila elements. In total, three sense and three anti-sense degenerate primers were selected. PCR amplification was performed as previously described (Glushkov *et al.*, 2006). 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. Sequences were deposited to GenBank under accessions DQ054404-054468, AY959213-959309.

Sequence analysis. Search of LTR elements in NCBI GenBank database (http://www.ncbi.nlm.nih.gov) was performed as described previously (Berezikov *et al*., 2000; Pruitt *et al.*, 2005). Selection of RT containing sequences was performed using a search of ORFs (ORF finder, http://www.ncbi.nlm.nih.gov/gorf/gorf.html) and comparison with GenBank database (Pruitt *et al*., 2005). The newly identified RT domain nucleotide sequences were analyzed in 3 steps: 1) aligning sequences using 4 different algorithms implemented in ClustalW, MAFFT, ProbCons, and MUSCLE programs; 2) correction of multiple sequence alignments using RASCAL software package, 3) reconstruction phylogenetic trees by maximum likelihood method using PHYML program (Guindon, Gascuel, 2003; Thompson *et al.*, 2003; Edgar, 2004; Do *et al.*, 2005; Katoh *et al.*, 2005).

IMPLEMENTATION AND RESULTS

A PCR survey of Ty1/copia and Ty3/gypsy retrotransposons. 19 different plant species covered 17 main plant taxa we tested by amplifying genomic DNAs with the selected degenerate oligonucleotides primers. The PCR results were considered positive if a fragment of the expected size was observed. Ty3/gypsy elements have been found in all species investigated. Ty1/copia elements have not been found only in *Marchantia polymorpha*. In contrast, Athila elements were tested only in several plant taxa (Fig. 1*a*). In order to confirm the results of PCR analysis, the PCR products were cloned and sequenced. Totally, 169 clones with reverse transcriptase sequences were determined: 66 sequences for Ty1/copia retroelements; 81 sequences for Ty3/gypsy retroelements; and 22 sequences for Athila retroelements.

Phylogenetic analysis. We conducted a phylogenetic analysis using the maximum likelihood method implemented in PHYML program combined with comparative analysis of all reconstructed trees by TreeJuxtaposer software (Fig. 1) (Munzner *et al.*, 2003). There can be distinguished three groups of elements: (i) Ty1/copia; and (ii) Ty3/gypsy, and (iii) Athila. All three groups of the elements formed separate branches with the high bootstrap values (100 %). For the detailed phylogenetic analysis Ty3/gypsy and Ty1/copia elements have been tested separately. All Ty1/copia elements form a monophyletic branch on the phylogenetic tree. The representatives of the most ancient plant taxa (except of Filicopsida) are close to the root of the tree and form separate branches.

a

	Copia	Ty3	Athila
Marchantiophyta			
1. Marchantia polymorpha	-	+	-
Bryophyta			
2. Physcomitrella patens	+	+	-
3. Fontinalis	+	+	-
Lycopodiophyta			
4.Lycopodium lagopus	+	+	-
Filicophyta			
5.Pteridium aquilinum	+	+	-
Equisetophyta			
6.Equisetum arvense	+	+	-
Cycadophyta			
7.Cycas revoluta	+	+	+
Ginkgophyta			
8.Ginkgo biloba	+	+	+
Gnetophyta			
9. Ephedra distachya	+	+	+
Cupressaceae			
10. Juniperus communis	+	+	-
Sciadopityaceae			
11. Sciadopitys verticillata	+	+	-
Taxaceae			
12. Taxus baccata	+	+	-
Pinaceae			
13. Pinus radiata	+	+	+
Nymphaeales			
14. Nymphaea tetragona	+	+	+
15. Brasenia schreberi	+	+	-
Magnoliales			
16. Magnolia grandiflora	+	+	-
Winterales			
17. Drimys buxifolia	+	+	-
Piperales			
18. Peperomia caperata	+	+	+
Rosids			
19. Pelargonium zonale	+	+	+

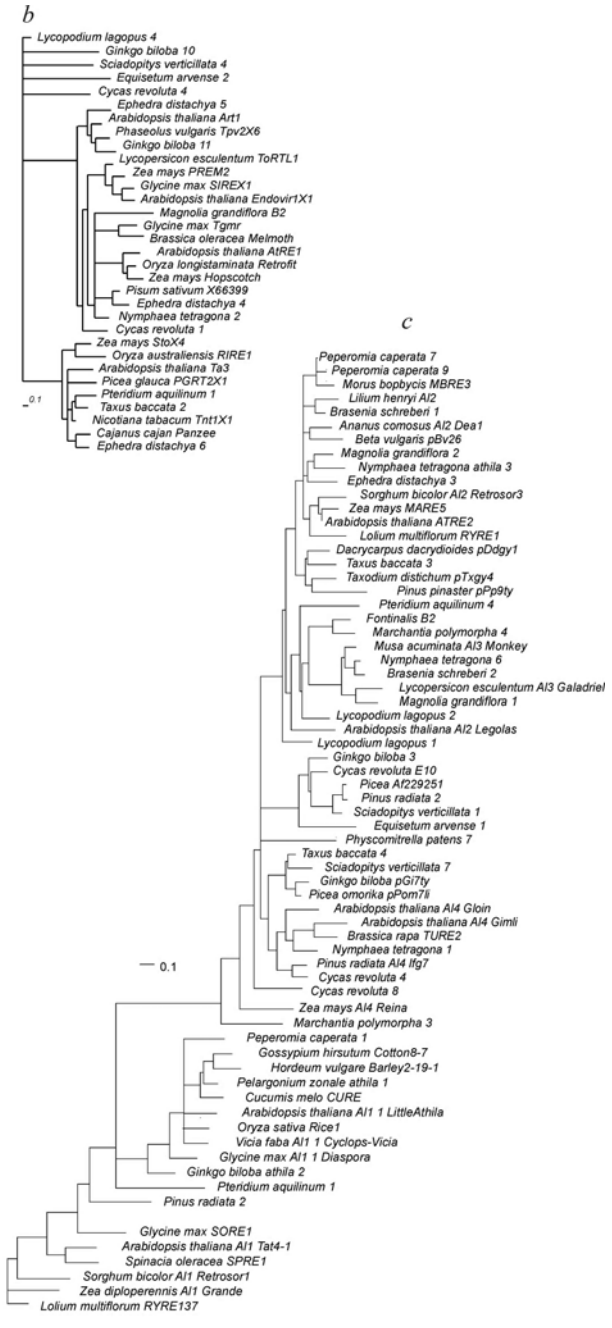


Figure 1. Distribution of three LTR retrotransposon groups among 19 plant genomes investigated (*a*); consensus tree of copia (*b*) and ty3 LTR retroelements (*c*).

At the same time, the rest of the clusters contain two and more species from the evolutionary distinct taxa that led us to a suggestion that there is more than one family of Ty1/copia elements in these plants. The rest of the examined elements are grouped in several phylogenetic branches, each of them contained the representatives of two and more different taxa. Phylogenetic analysis of Ty3/gypsy elements was conducted taking into account previously described groups of Ty3/gypsy elements. Three large groups were distinguished: Athila, Tat and Chromovirus (Marin, Llorens, 2000). On the common nucleotide tree Ty3/gypsy elements, obtained in the present study, are clustered in two groups: Athila and Chromoviruses. Athila group contains 13 elements from five different

species (*Ephedra distachya*, *Ginkgo biloba*, *Pinus radiata*, *Peperomia caperata*, *Pelargonia zonale*), which are grouped in the species-specific branches clearly separated each other. As in the case of Ty1/copia elements, several phylogenetic clusters have been found in Chromoviruses. Within these clusters the elements are grouped in the species-specific manner. This confirms a suggestion that Chromoviruses contain several separate families of LTR retrotransposons.

DISCUSSION

Both Ty3/gypsy and Ty1/copia elements have been found in all plant species tested. These results are in a good agreement with the known data on the LTR retrotransposon distribution that place an origin of these transposable elements before the plant origin. Representatives of the Athila group of Ty3/gypsy elements have not been found in the ancient plant taxa (Bryophyta, Equisetopsida, Filicopsida, and Lycopsida). Origin of these elements is connected with an appearance of the seed plants. Several new families have evolved in both Ty3/gypsy and Ty1/copia groups.

REFERENCES

Berezikov E., Bucheton A., Busseau I. (2000) A search for reverse transcriptase-coding sequences reveals new non-LTR retrotransposons in the genome of Drosophila melanogaster. *Genome Biol*., **1**(6), RESEARCH0012.

Do C.B., Mahabhashyam M.S.P., Brudno M., Batzoglou S. (2005) PROBCONS: probabilistic consistency-based multiple sequence alignment. *Genome Res.*, **15**(2), 330–340.

Edgar R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**(1), 113.

Glushkov S., Novikova O., Fet V., Blinov A. (2006) Divergent non-LTR retrotransposon lineages from the genomes of scorpions (Arachnida: Scorpiones). *Mol. Genet. Genomics*, **275**, 288–296.

Guindon S., Gascuel O. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol*., **52**(5), 696–704.

Kidwell M.G., Lisch D.R. (2000) Transposable elements and host genome evolution. *TREE*, **15**(3), 95–98.

Marin I., Llorens C. (2000) Ty3/Gypsy retrotransposons: description of new Arabidopsis thaliana elements ans evolutionary perspectives derived from comparative genomic data. *Mol. Biol. Evol*., **17**, 1040–1049.

Munzner T., Guimbretiere F., Tasiran S., Zhang L., Zhou Y. (2003) TreeJuxtaposer: scalable tree comparison using Focus + Context with guaranteed visibility. *ACM Transactions on Graphics*, **22**(3), 453–462.

Pruitt K.D., Tatusova, T., Maglott D.R. (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucl. Acids Res*., **33**(Database issue):D501–D504.

Rogers S.O., Bendich A.J. (1985) Extraction of DNA from milligram amounts of fresh, herbarium and mummified plant tissues. *Plant Mol. Biol*., **5**, 69–76.

Rose T.M., Schultz, E.R., Henikoff, J.G., Pietrokovski, S., McCallum, C.M., Henikoff, S. (1998) Consensus-degenerate hybrid oligonucleotide primers for amplification of distantly related sequences. Nucleic Acids Res., 26: 1628-1635.

Thompson J.D., Thierry J.C., Poch O. (2003) RASCAL: rapid scanning and correction of multiple sequence alignments. *Bioinformatics*, **19**(9), 1155–1161.

NEW KIDS ON THE BLOCK: SELF-SYNTHESIZING DNA TRANSPOSONS

Kapitonov V.V.*[*], *Jurka J.
Genetic Information Research Institute, Mountain View, California, USA
[*] Corresponding author: e-mail: vladimir@girinst.org

Key words: transposons, molecular evolution, genomics, DNA polymerase, integrase, cysteine protease, ATPase, computational biology

SUMMARY

Motivation: Transposable elements constitute quite significant static and dynamic components of eukaryotic genomes. Moreover, transposable elements serve as efficient tools in genetic engineering. Therefore, identification and studies of transposable elements are important for researchers working in different fields of molecular biology.

Results: Based on computational studies of eukaryotic genomes, we discovered a novel class of DNA transposons, called *Polintons*, characterized by a unique set of proteins including a protein-primed family B DNA polymerase, retroviral integrase, cysteine protease, and ATPase. *Polintons* are also characterized by 6-bp target site duplications, long terminal inverted repeats, and 5′-AG and TC-3′ termini. According to a transposition model discussed here, a *Polinton* DNA molecule excised from the genome by the *Polinton*-encoded integrase serves as a template for extrachromosomal synthesis of its double-stranded DNA copy by the *Polinton*-encoded DNA polymerase and is inserted into genome by the integrase.

Availability: http://www.girinst.org/repbase/.

INTRODUCTION

Genomes of most eukaryotes are populated by DNA copies of parasitic elements known as transposable elements (TEs) capable of reproducing themselves in the host genome in a non-Mendelian fashion. Despite an enormous diversity of eukaryotic TEs, they belong to only two types called retrotransposons and DNA transposons (Craig *et al*., 2002). While a retrotransposon is transposed via reverse transcription of its mRNAs, a DNA transposon is transposed via transfer of its genomic copy from one genomic site to another. Transposition of a retrotransposon is catalyzed by reverse transcriptase (RT) and endonuclease (EN) domains of a polyprotein encoded by itself or by other retrotransposons. All retrotransposons can be further divided into two subclasses called LTR and non-LTR retrotransposons. An mRNA molecule expressed during transcription of the genomic non-LTR retrotransposon is reverse transcribed and inserted in the genome. An LTR retrotransposon may carry three open reading frames (ORFs) coding for the *gag*, *env*, and *pol* proteins, the latter is composed of the RT, EN, and aspartyl protease domains. The endonuclease domain in LTR retrotransposons is usually called integrase (INT) and is distantly related to the DDE transposases (TPase) encoded by *Mariner* DNA transposons (Capy *et al*., 1997).

DNA transposons identified so far in eukaryotes belong to two classes characterized by the so-called “cut and paste” (Craig, 1995) and “rolling-circle” (Kapitonov, Jurka, 2001) mechanisms of transposition. Unlike retrotransposons, which synthesize their DNA

copies using their own RNA-dependent DNA polymerase (RT), DNA transposons cannot synthesize DNA. Instead, they multiply using the host replication machinery (Kapitonov *et al*., 2004). A typical autonomous *mariner*, *hAT*, *piggyBac*, *P*, *Merlin*, *Transib* DNA transposon encodes a single protein called transposase, which acts as an endonuclease and catalyses transfer of transposon DNA strands from one genomic site to another (Craig *et al*., 2002; Kapitonov, Jurka, 2003; Feschotte, 2004; Kapitonov, Jurka, 2005). In the *En/Spm*, *MuDR*, *Harbinger*, and *Helitron* superfamilies, an autonomous transposon usually encodes TPase and DNA-binding proteins (Kapitonov, Jurka, 2001; Craig *et al*., 2002; Kapitonov, Jurka, 2004).

RESULTS AND DISCUSSION

Polintons are widespread in protists, fungi and animals including entamoeba, trichomonas, soybean rust, sea urchin, sea anemone, sea squirt, fishes, chicken, lizard, frog, insects, and worms (Kapitonov, Jurka, 2006). Autonomous *Polintons* are 10–20 kb long and encode up to ten different proteins including a family B DNA polymerase (POLB), retroviral-like integrase, adenoviral-like protease (PRO), and putative ATPase (ATP). Polintons are the most complex DNA transposons in eukaryotes. Based on structural and evolutionary characteristics of these transposons, we developed a model of *Polinton* transposition. We also discuss implications of our findings, including likely origin of *Polintons* from a linear plasmid and evolution of adenoviruses from an ancient *Polinton*.

Given known distant similarities between retroviral INTs and "cut and paste" TPases, one can expect "cut and paste" transpositions of *Polintons* catalyzed by their INT. However, arguments listed below strongly suggest that transposition of *Polintons* follows a completely different mechanism unseen previously in transposons. First, a perfect conservation of all functional motifs in the extremely diverged POLBs indicates that the DNA-DNA polymerase and proofreading activities are necessary for transposition of *Polintons*. Second, POLB in *Polintons* belongs to the group of protein-primed DNA polymerases encoded by genomes of bacteriophages, linear plasmids and adenoviruses. Third, all these genomes and *Polintons* are characterized by terminal inverted repeats that are usually several hundred bps long. Fourth, their termini are composed of short 1-3-bp tandem repeats, which are thought to be necessary for the slide-back mechanism in protein-primed DNA synthesis (Mendez *et al*., 1992).

Table 1. General properties of Polintons in different species

Species	Copies per genome	Encoded proteins
Frog, Zebrafish	~100	**POLB**, **INT**, PRO, ATP, PW, PX, PY, PZ
Lancelet, Sea squirt/urchin	50–100	**POLB**, **INT**, PRO, ATP, PW, PX, PY, PZ
Red flour beetle	~50	**POLB**, **INT**, PRO, ATP, PW, PX, PY, PZ
Nematode	~100	**POLB**, **INT**, PRO, ATP, PY, PC1, PC2
Nematostella vectensis	~100	**POLB, INT,** PRO, ATP, PX, PY, PZ
Trichomonas vaginalis	~1000	**POLB**, **INT**, ATP, ATP1, PTV1-PTV6
Entamoeba invadens	50–100	**POLB, INT,** ATP, ATP1, PTV6

We proposed that *Polintons* form a novel class of DNA transposons propagated through protein-primed self-synthesis by POLB, according to the next model. First, during host genome replication, the integrase-catalyzed excision of *Polinton* from the host DNA leads to an extrachromosomal single-stranded *Polinton* that forms a racket-like structure. Second, the *Polinton* POLB replicates the extrachromosomal *Polinton*. Given the arguments listed above, initiation of the replication requires the terminal protein (TP) that binds a free 5′ end of *Polinton*. It is thought that N-terminal domains of proteins, whose C-terminal parts serve as POLB, encode TP in some linear plasmids. Therefore, it

is likely the N-terminal 400–600-aa domain of the *Polinton* POLB serves also as TP. After the double-stranded *Polinton* is synthesized, the INT molecules bind its termini and catalyze its integration in the host genome.

Polintons are present in genomes of species that belong to diverse eukaryotic kingdoms including opisthokonts (metazoa and fungi), heterokonts (oomycetes), alveolates (ciliates), amoebozoa (entamoeba), and parabasalids. Given the conserved complex structure of *Polintons*, their monophyletic origin is most likely. Although *Polintons* are much more complex (up to eight conserved proteins) than known eukaryotic TEs and resemble viruses (adenoviruses and BmDNV-2), we did not find any *Polinton* protein similar to viral capsid or envelope proteins, which are necessary for the infectious transmission of viruses. Moreover, we are not aware of any viruses capable of spreading over different kingdoms. Most likely, *Polintons* emerged in a common ancestor of modern species from the eukaryotic crown, approximately a billion years ago. As we reported here, *Polintons* share their main structural characteristics with "selfish" linear plasmids, bacteriophages and adenoviruses that multiply using their protein-primed DNA polymerases. Linear plasmids can be split into two groups: (i), plasmids that exist in mitochondria of plants and fungi; (ii), plasmids that exist in the yeast cytoplasm. While it is likely that mitochondrial linear plasmids evolved from bacteriophages during the evolution of mitochondria from bacteria, different equally plausible scenarios puzzle understanding of the evolution of cytoplasmic plasmids. Although *Polintons* represent a previously unknown link between cytoplasmic plasmids/adenoviruses and mitochondrial plasmids/bacteriophages, many aspects of evolution of *Polintons* and cytoplasmic linear plasmids remain unclear. Acquisition of the integrase by a protein-primed replicating genome of an ancient virus or linear plasmid was the most certain stage of the evolution. The *Polinton* INT has evolved from an INT encoded by an LTR retrotransposon. Thus, it might have been acquired after integration of an ancient LTR retrotransposon into the ancestral linear genome. However, we cannot rule out the origin of the *Polinton* INT from a DNA transposon. For instance, the *Tdd-4* transposon from the slime mould *Dictyostelium discoideum* genome is a DNA transposon characterized by its 145-bp TIRs, 5-bp TSDs, and a TPase that is similar to INTs encoded by LTR retrotransposons (Wells, 1999).

Polintons are characterized by a highly patchy distribution in different species. In insects, *Polintons* are present in flies and beetles but absent in mosquitoes and bees. In fungi, they are present in basidiomycetes (soybean rust) and glomeromycetes (*G. intraradices*) but absent in ascomycetes (including the completely sequenced yeast genome). We interpret this patchiness as a frequent loss of *Polintons* from genomes. Due to the high complexity of *Polintons*, their transposition may be tightly regulated and may explain their small numbers in most studied genomes.

ACKNOWLEDGEMENTS

This work was supported by the National Institutes of Health grant 5 P41 LM006252–08.

REFERENCES

Capy P. *et al*. (1997) Do the integrases of LTR-retrotransposons and class II element transposases have a common ancestor? *Genetica*, **100,** 63–72.

Craig N.L. (1995) Unity in transposition reactions. *Science*, **270,** 253–254.

Craig N.L. *et al*. (eds.) (2002) *Mobile DNA II* (ASM Press, Washington, DC).

Feschotte C. (2004) Merlin, a new superfamily of DNA transposons identified in diverse animal genomes and related to bacterial IS1016 insertion sequences. *Mol. Biol. Evol.*, **21,** 1769–1780.

Kapitonov V.V., Jurka J. (2001) Rolling-circle transposons in eukaryotes. *Proc. Natl. Acad. Sci. USA*, **98,** 8714–8719.

Kapitonov V.V., Jurka J. (2003) Molecular paleontology of transposable elements in the *Drosophila melanogaster* genome. *Proc. Natl. Acad. Sci. USA*, **100,** 6569–6574.

Kapitonov V.V., Jurka J. (2004) Harbinger transposons and an ancient HARBI1 gene derived from a transposase. *DNA Cell Biol.*, **23,** 311–324.

Kapitonov V.V. *et al*. (2004) Anthology of human repetitive DNA. In Meyers R.A. (ed) *Encyclopedia of Molecular Cell Biology and Molecular Medicine*. Wiley-VCH, Weinheim, Germany, pp. 251–305.

Kapitonov V.V., Jurka J. (2005) RAG1 core and V(D)J recombination signal sequences were derived from Transib transposons. *PLoS Biol.*, **3,** e181, 998–1011.

Kapitonov V.V., Jurka J. (2006) Self-synthesizing DNA transposons in eukaryotes. *Proc. Natl. Acad. Sci. USA*, **103,** 4540–4545.

Mendez J. *et al*. (1992) Initiation of phi 29 DNA replication occurs at the second 3' nucleotide of the linear template: a sliding-back mechanism for protein-primed DNA replication. *Proc. Natl. Acad. Sci. USA*, **89,** 9579–9583

Wells D.J. (1999) Tdd-4, a DNA transposon of Dictyostelium that encodes proteins similar to LTR retroelement integrases. *Nucl. Acids Res.*, **27,** 2408–2415.

Multi-SNP ANALYSIS OF CCR5-CCR2 GENES IN ETHIOPIAN JEWS: MICRO-EVOLUTION AND HIV-RESISTANCE IMPLICATIONS

***Korostishevsky M.*[*1], *Bonne'-Tamir B.*[1], *Bentwich Z.*[2], *Tsimanis A.*[2]**
[1] Department of Human Molecular Genetics and Biochemistry, Tel Aviv University, Israel; [2] Institute of Clinical Immunology and AIDS, Kaplan Medical Center, Rehovot, Israel
[*] Corresponding author: e-mail: korost@post.tau.ac.il

Key words: multi-SNP analysis, genetic trees, CCR5-CCR2 genes, HIV resistance

SUMMARY

Motivation: This is the first report concerning CR5 SNPs in the Ethiopian Jewish population. CCR5 and CCR2 genes have been implicated in HIV disease progression, resistance or non-progressive infection. To determine the influence of host genetics on HIV infection, we examined 29 HIV-seronegative individuals of Ethiopian descent for polymorphisms in the CCR5-CCR2 gene region and compared the results with those of 13 exposed but uninfected individuals. Multi-SNP analysis was used for sample comparisons and for population relationship estimates.

Results: Using multi-SNP analysis, no significant differences in the genotypes frequencies between the studied groups was found ($\chi^2_{3df} = 4.662$, $p = 0.198$). The pattern of CCR5-CCR2 genetic variations in Ethiopian Jews resembles the one found in Asian populations and is distinguished from the one found in Africans.

Availability: http://www.interscience.wiley.com/jpages/1552-4841/suppmat/ajmg.b.30212.html.

INTRODUCTION

The chemokine receptor CCR5 is an essential co-receptor for the cellular entry of R5 strains of HIV-1, which predominate in the early stages of infection (Moore *et al*., 1997). Following infection with HIV-1, the majority of patients develop AIDS within 10 years, a small subset of infected individuals rapidly progress to AIDS, and less than 5 % – remain asymptomatic without antiretroviral therapy. There is evident that, even after being repeatedly exposed to HIV, some individuals remain seronegative.

It has been proposed that HIV infection and disease progression might be genetically controlled. Functionally important polymorphisms in the regulatory region of CCR5 gene, a 32-base pair deletion in the coding part of CCR5 (CCR5-Δ32) and a single conservative valine-to-isoleucine (V64I) mutation in CCR2 coding region (CCR2-G190A) have been identified (Carrington *et al*., 1999). These polymorphisms are distributed through human populations with differing frequencies depending on the ethnic groups or on particular population groups that contain distinctive set of haplotype pair combinations (Gonzalez *et al*., 1999). The finding that polymorphisms in the promoter region of CCR5 are associated with differential HIV-1 disease progression and susceptibility of cells to HIV-1 infection suggests that these haplotypes are not functionally similar.

The best genetic feature characterized is the CCR5-Δ32 deletion that results in synthesis of a short, nonfunctional CCR5 protein and the absence of cell surface CCR5

expression. Thus, the mechanism of protection most likely involves a reduction in the number of CCR5-positive target cells. The CCR2-V64I mutation confers resistance to AIDS progression, probably due to the heterodimerization and sequestration of the CCR5 receptor (Mellado *et al*., 1999). Mutations in the CCR5 promoter region determine the level of CCR5 gene transcription and production of the corresponding mRNA. It has also been hypothesized that polymorphisms in the CCR5 promoter region may influence cell surface expression and consequently could influence individual susceptibility to HIV. However, these data must be used with caution: firstly, a recent cohort study of Ugandan population shows no association between CCR5 polymorphisms and the rate of disease progression (Ramaley *et al*., 2002). Secondly, similar expression level of CCR5 was found in HIV-exposed uninfected female prostitutes and in unexposed control individuals from Kenya and Ethiopia (Fowke *et al*., 1998; Messele *et al*., 2001). Thirdly, *in vitro* infection study revealed that PBMC isolated from HIV-highly exposed uninfected and unexposed Thai women carrying different CCR5 haplogroups, had no differences in susceptibility to HIV-1 infection (Kulkarni *et al*., 2003).

Present-day Ethiopian Jews lived in the north of Lake Tana in Gondar. During Ethiopian civil war (1984–1985) and then in 1989, several thousands of Ethiopian Jews were airlifted to Israel. Some of them arrived in Israel with dramatic health problems including a variety of immune-mediated and infectious diseases. There is abundant evidence that chemokines and cytokines production, which is under genetic control, may exert certain disease occurrences and outcomes.

Using evolutionary-based CCR5 haplotype classification (Gonzalez *et al*., 1999), we have characterized the DNA polymorphisms at the loci that encode CCR5 and CCR2 receptors, in two groups of Ethiopian Jews: healthy individuals without any history of HIV infection and individuals who were exposed but uninfected. We verified the presence of CCR5-Δ32 deletion and genotyped sequence variations for the single-nucleotide polymorphism (SNP) G208T, T627C, A676G and C927T in the CCR5 promoter region as well as G190A mutation in the coding part of CCR2 gene.

We estimated the magnitude of LD between the SNPs and performed multi-SNP analysis between the samples. Using the haplotype distribution data for different ethnic groups, we investigated the phylogenetic relationships for Ethiopian Jews.

SAMPLES AND METHODS

A total of 29 HIV-1-negative (control) and 13 exposed but uninfected seronegative (ESN) Ethiopian Jews were sampled for this study. The evaluation of the clinical status of individuals including the presence of antibodies to HIV-1 and HIV-1 viral particles has been performed at the Kaplan Medical Center, Rehovot, Israel. All samples were coded and blind-tested. Informed consent was obtained for the collected samples.

We used genomic DNA obtained from peripheral blood lymphocytes. The DNA samples were subjected to a polymerase chain reaction-restriction fragment length polymorphism (PCR-RFLP) assay. PCR amplification was performed to amplify CCR5 promoter and CCR5 and CCR2 genes fragments covering the polymorphic sites.

Possible differences in the frequency of each of the SNP genotypes or alleles between the samples were estimated using the $\chi 2$ test, as described elsewhere. The Arlequin software package, http://Lgb.unige.ch/arlequin, was used to evaluate genetic distances between different populations, and to calculate the maximum likelihood (ML) of haplotype frequencies. Based on the ML haplotype frequency estimates, the likelihood ratio test (LRT) for sample differentiation was performed as we previously described (Korostishevsky *et al*., 2006). PHYLIP (http://evolution.gs.washington.edu/phylip.html) package was used for phylogeny inferences based on the CCR2-CCR5 region genetic distances.

RESULTS AND DISCUSSION

Two groups of Ethiopian Jews were genotyped for the CCR5-Δ32, CCR2-G190A and CCR5 promoter alleles constituting the CCR5 human haplotypes. We performed DNA PCR by use of primers that amplified the region encoding the 32-bp deletion. No CCR5-Δ32 deletions were detected in both groups. CCR5-Δ32 allele is very common in Caucasians, but no such allele was reported in people of African and Asian descent including Ethiopian Jews (Kantor, Gershoni, 1999).

A discrepancy between the control and ESN individuals was observed for the T627C SNP only ($\chi 2_{1df} = 4.140$, $p = 0.042$). This deviation in allele frequency of one from 5 SNPs was not preserved after the Bonferroni correction (data not shown). The haplotype distributions in the samples are presented Table 1. The multi-SNP likelihood ratio test (Korostishevsky *et al*., 2006) did not elicit significant differences between the studied groups ($\chi^2_{df3} = 4.66$, $p = 0.198$).

Table 1. ML estimates of haplotype frequencies in ESN and Control samples

Haplotype	HH-code	ESN	Control	OR	LRT* (p-value)
G-G-C-A-C	HHE	0.385	0.207	1.859	4.66 (0.198)
G-T-T-G-C	HHC	0.231	0.431	0.535	
A-G-C-A-T	HHF*2	0.269	0.207	1.301	
G-G-T-A-C	HHA	0.115	0.155	0.744	

Our results indicate that the HHF*2 frequency is slightly higher in the ESN individuals than in the group of HIV-1-negative Ethiopian Jews (26.9 % *vs.* 20.7 %), although the difference did not attain statistical significance. In the control group, the most common CCR5 haplotype was HHC (43.1 %), but in the group of ESN individuals the most common was HHE (38.6 %). These two haplotypes have significantly higher frequencies in Caucasians (Gonzalez *et al*., 2000). The minor haplotype in both studied groups was HHA (15.5 % and 11.5 %, respectively), which was more frequent in Africans. Genotyping of 29 HIV-negative and 13 ESN Ethiopian Jews failed to detect presence of HHD and/or HHB haplotypes, which were reported as specific to African population.

Recently, a comparative analysis of CCR5 polymorphisms in HIV-exposed uninfected individuals from two ethnic groups, Caucasian and Asian, was undertaken (Gonzalez *et al*., 1999; Mangano *et al*., 2000). Nevertheless, *in vitro* infection experiments showed that PBMC isolated from the HIV-1-exposed and unexposed seronegative women carrying different CCR5 haplogroup had no differences in susceptibility to HIV-1 infection (Yang *et al.*, 2003).

The genetic affinities of the Ethiopian Jews were investigated using classical autosomal markers, as well as DNA, mtDNA and Y-chromosome markers. It was demonstrated that the Ethiopian Jews are a mixture of African and Caucasian (Asian) population and are significantly different from other Jewish communities. Several authors argued that Ethiopian Jews were derived mostly from Africans. However, both cultural and historic evidence suggest close affinity between Ethiopians and Asian populations (Near East and southern Arabia) (Ritte *et al*., 1993; Hammer *et al*., 2000).

We used a tree reconstruction to investigate population relationships according to CCR5-CCR2 polymorphism. The dendrogram shows two well-defined groups. The first one contains two populations: Non-Pygmy and American-Africans. The second group contains the remaining six populations which are further divided. Within this group, Ethiopian Jews are “sisters” to Asian populations (non-Indians and Indians), while the Thai population is found to be the most distant (Fig. 1).

In conclusion, our results indicate that there is no significant difference in the distribution of CCR5 haplotypes among ESN and control individuals. We also did not observe preponderance of AIDS-“protective” 303-G, 627-T and 676-G alleles in ESN

individuals. On the contrary, frequencies of these alleles are higher in the control group. The pattern of CCR5-CCR2 genetic variations in Ethiopian Jews resembles the one found in Asian populations and is distinguished from that found in Africans.

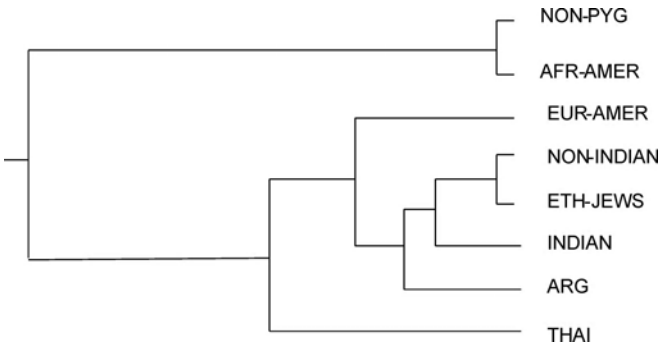


Figure 1. UPGMA tree based on CCR5-CCR2 haplotype frequencies.

REFERENCES

Carrington M. *et al*. (1999) Genetics of HIV-1 infection: chemokine receptor CCR5 polymorphism and its consequences. *Hum. Mol. Genet.*, **8**, 1939–1945.

Fowke K.R. *et al*. (1998) HIV type 1 resistance in Kenyan sex workers is not associated with altered cellular susceptibility to HIV type 1 infection or enhanced chemokine production. *AIDS Res. Hum. Retroviruses*, **14**, 1521–1530.

Gonzalez E. *et al*. (1999) Race-specific HIV-1 disease-modifying effects associated with CCR5 haplotypes. *PNAS*, **96**, 12004–12009.

Hammer M.F. *et al*. (2000) Jewish and Middle Eastern non-Jewish populations share a common pool of Y-chromosome biallelic haplotypes. *PNAS*, **97,** 6769–6774.

Kantor R., Gershoni J.M. (1999) Distribution of the CCR5 gene 32-base pair deletion in Israeli ethnic groups. *JAIDS*, **20**, 81–84.

Korostishevsky M. *et al*. (2006) Transmission disequilibrium and haplotype analyses of the G72/G30 locus: Suggestive linkage to schizophrenia in Palestinian Arabs living in the North of Israel. *Amer. J. Med. Genet.*, **141**B, 91–95.

Kulkarni P.S. *et al*. (2003) Resistance to HIV-1 infection: lessons learned from studies of highly exposed persistently seronegative (HEPS) individuals. *AIDS Rev*., **5**(2), 87–103.

Mangano A. *et al*. (2000) Concordance between the CC chemokine receptor 5 genetic determinants that alter risks of transmission and disease progression in children exposed perinatally to human immunodeficiency virus. *J. Infect. Dis*, **183**,1574–1585.

Mellado M. *et al*. (1999) Chemokine control of HIV-1 infection. *Nature*, **400**, 723–724.

Messele T. *et al*. (2001) No difference in *in vitro* susceptibility to HIV type 1 between high-risk HIV-negative Ethiopian commercial sex workers and low-risk control subjects. *AIDS Res. Hum. Retroviruses*, **17**, 433–441.

Moore J.P. *et al*. (1997) Co-receptors for HIV-1 entry. *Curr Opin Immunol.*, **9**, 551–562.

Ramaley P.A. *et al*. (2002) Chemokine-receptor genes and AIDS risk. *Nature*, **417**, 140.

Ritte U. *et al*. (1993) The differences among Jewish communities-maternal and paternal contributions. *J. Mol. Evol.*, **37,** 435–440.

Yang C. *et al*. (2003) Polymorphisms in the CCR5 coding and noncoding regions among HIV type 1-exposed, persistently seronegative female sex-workers from Thailand. *AIDS Res. Hum. Retroviruses*, **19**, 661–665.

TRANSCRIPTIONAL REGULATION OF THE METHIONINE BIOSYNTHESIS IN ACTINOBACTERIA AND STREPTOCOCCI

Kovaleva G.Yu. [*1, 2], ***Gelfand M.S.*** [1, 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

Key words: comparative genomics, methionine biosynthesis, transcriptional regulation

SUMMARY

Motivation: An evolutionary scenario has been proposed for the regulatory mechanisms of methionine biosynthesis in gram-positive bacteria based on the distribution of T- and S-boxes in various genomes. However, some bacteria, such as *Corynebacterium glutamicum* and *Streptococcus agalacticae*, use regulatory DNA-binding proteins to control methionine biosynthesis genes instead of these RNA-based mechanisms. In this study we focused on evolution of these transcription factors in *C. glutamicum*, *S. agalacticae* and related species.

Results: We performed detailed analysis of the orthologous regulons for transcription factors regulating methionine biosynthesis in related genomes. We predicted two new potential members of the methionine biosynthesis pathway in several genomes of corynebacteria, and identified potential binding signals for methionine transcription factors in some relatively distant and diverse genomes, such as *Bifidobacterium* and *Brevibacterium*.

INTRODUCTION

Methionine is an essential amino acids and the universal N-terminal amino acid of proteins, and because of that biosynthesis of methionine is extensively studied in various organisms to allow for their future application in biotechnological production of the methionine. Accumulating experimental data allowed us to describe a potential scenario for evolution of methionine biosynthesis regulation in bacterial genomes (Rodionov *et al*., 2004). As most gram-positive bacteria use T- and S-box RNA-based mechanisms of methionine biosynthesis regulation, this scenario mostly relies on the distribution of T- and S-boxes in various genomes.

Nevertheless, some gram-positive bacteria regulate the methionine biosynthesis pathway by DNA-binding regulatory proteins. These bacteria, among others, include *Corynebacterium glutamicum* and *Strepcococcus agalacticae* species, that are the focus of this study. Transcriptional regulation of the methionine biosynthesis in the *Corynebacterium glutamicum* genome is well studied (Rey *et al*., 2005). The regulatory protein is McbR (a member of the TetR protein family) with the binding site consensus TAGAC-N6-GTCTA. The operon structures for members of the McbR regulon also were predicted.

The transcriptional regulator of the methionine biosynthesis in *Streptococcus agalacticae* is MtaR, a member of the LysR protein family (Shelver *et al*., 2003). The consensus signal for regulators of the LysR family is T-N_{11}-A (Schell, 1993). The

predicted binding signal for the methionine regulatory protein in streptococci is TATAGTTtnaAACTATA (Rodionov *et al*., 2004). As no binding sites of MtaR have been characterized in experiment, it is not clear whether this is the signal of MtaR.

In this study we analyzed the evolution of the candidate MtaR and McbR regulons in an attempt to extend the evolutionary scenario for methionine regulatory mechanisms.

METHODS AND ALGORITHMS

Complete and partial sequences of bacterial genomes were extracted from GenBank (Benson *et al*., 1999).

For identification of orthologous genes and site patterns, the Genome Explorer program (Mironov *et al*., 1999) was used. SignalX (Mironov *et al*., 2000) was used to construct nucleotide weight matrices. Multiple sequence alignments were done using ClustalX (Thompson *et al*., 1997). Phylogenetic trees of proteins were constructed by the maximum likelihood method implemented in PHYLIP (Felsenstein, 1981).

RESULTS AND DISCUSSION

Orthologs of McbR are present in four closely related genomes of genus *Corynebacterium*: *C.glutamicum*, *C.diphtheriae*, *C.efficiens* and *C.jeikeium*, and in three more distant genomes of other Actinomycetales: *Nocardia farcinica*, *Streptomyces coelicolor* and *Leifsonia xyli*. As the operon structure for members of the McbR regulon in *C.glutamicum* was already predicted (Rey *et al*., 2005), we focused on orthologous regulons in other genomes. The comparison of the operon structures reveals multiple positional rearrangements even in the closely related genomes. All observed rearrangements could be classified as follows:

1. Nonorthologous replacement (threonine synthase and alkansulfonate monooxygenase);
2. Multiple operon rearrangements (mostly rearragements of the gene order in operons);
3. Duplications creating paralogs;
4. Loss of orthologs (mostly in parasitic genomes);
5. Absolutely conserved operon structure was observed only for two of 22 operons, both are monocistronic.

Overall, even in closely related genomes of the genus *Corynebacterium,* we observed multiple and diverse evolutionary events despite early observations of high level of the genome stability in corynebacteria (Nakamura *et al*., 2003).

Our analysis suggest that methionine regulators of distant genomes of *Nocardia farcinica*, *Leifsonia xyli* and *Streptomyces coelicolor* use binding signal(s) that differ from McbR signal in corynebacteria. Therefore, we concentrate on potential binding sites for McbR only in related genomes of genus *Corynebacterium*. Our observations demonstrate conservation of regulation of the methionine biosynthesis genes in all these genomes. The combined data on the analysis of regulatory regions and positional comparison allowed us to predict two new candidate members of McbR regulon. One of them potentially encodes glutamine-amidotransferase and the only explanation of its functional role in methionine biosynthesis pathway is the regulation of synthesis on the level of aspartate formation. The functional role of the second potential member of McbR regulon is much clearer, as the corresponding protein contains the methylthioadenosine (MTA) nucleosidase domain. MTA nucleosidase catalyzes the first reaction of methionine pool restoration from MTA, the by-product of the polyamine biosynthesis pathway.

The phylogenetic tree for the MtaR protein of *S.agalacticae* (Fig. 1) shows that the closest homologs of MtaR form two branches (marked in Fig. 1). The branch contents

differ mainly by complementary distant genomes, such as *Synthromonas*, *Bifidobacterium*, *Brevibacterium*, *Clostridium* and *Lactobacillus* for one branch, and *Enterococcus* for other. Both branches also contain the MtaR homologs from *Lactococcus lactis*, one of which has been experimentally characterized as the methionine biosynthesis regulator FhuR with the binding signal TAAAWWWTTTTA (Sperandio *et al*., 2005).

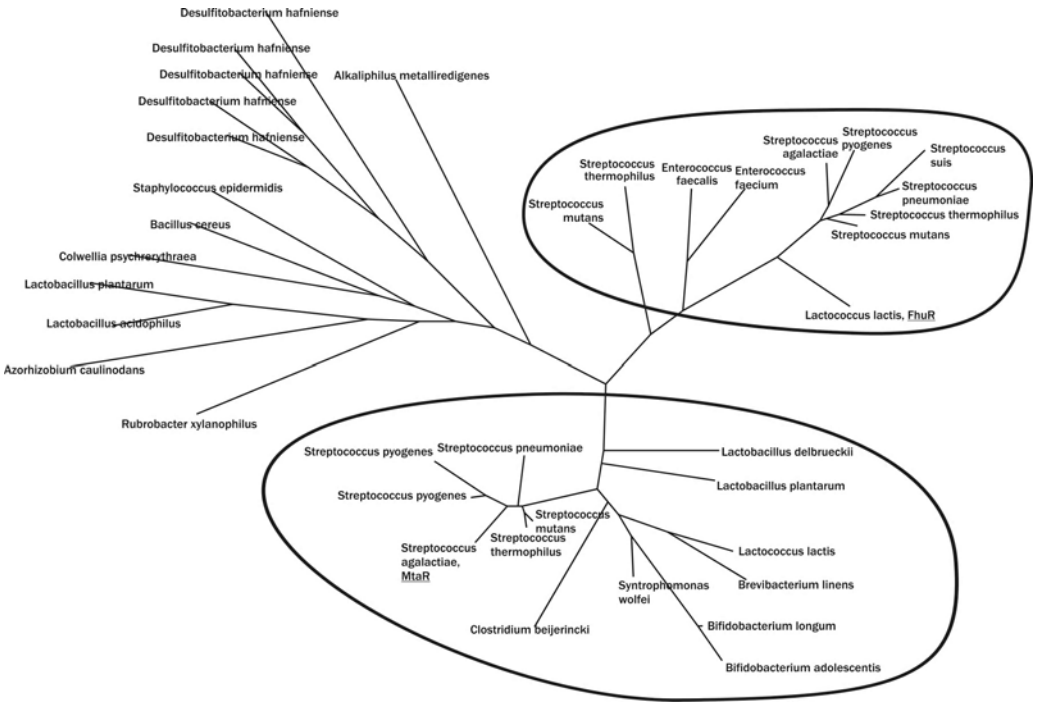


Figure 1. Phylogenetic tree of *S. agalacticae* MtaR and its relatives.

Overall, both branches contain regulators that potentially recognize different binding signals, ascribing the predicted streptococcal signal (Rodionov *et al*., 2004) to the MtaR protein of *S. agalacticae*. To verify this assumption we searched for binding signal(s) in the genomes that contain the only one of two paralogous regulators, such as *Bifidobacterium*, *Brevibacterium*, *Enterococcus* etc. Our initial observations seem to confirm the potential subdivision of specific binding signal(s) into two groups corresponding to the regulators subdivision in the phylogenetic tree. If our assumption will be proven in the following analysis, we could describe consecutive evolutionary events as methionine regulator duplication in the genome of common ancestor of streptococci followed by horizontal transfer of one of the paralogs into distant genomes.

ACKNOWLEDGEMENTS

This study was partially supported by the Howard Hughes Medical Institute and the Russian Academy of Sciences (program “Molecular and Cellular Biology”).

REFERENCES

Benson D.A., Boguski M.S., Lipman D.J., Ostell J., Ouellette B.F., Rapp B.A., Wheeler D.L. (1999) GenBank. *Nucl. Acids Res*., **27**, 12–17.

Felsenstein J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, **17**, 368–376.

Mironov A.A., Koonin E.V., Roytberg M.A., Gelfand M.S. (1999) Computer analysis of transcription regulatory patterns in completely sequenced bacterial genomes. *Nucl. Acids Res.*, 27, 2981–2989.

Mironov A.A., Vinokurova N.P., Gelfand M.S. (2000) GenomeExplorer: software for analysis of complete bacterial genomes. *Mol. Biol*., **34**, 222–231.

Nakamura Y., Nishio Y., Ikeo K., Gojobori T. (2003) The genome stability in Corynebacterium species due to lack of the recombinational repair system. *Gene*, **317**, 149–155.

Rey D.A., Nentwich S.S., Koch D.J., Ruckert C., Puhler A., Tauch A., Kalinowski J. (2005) The McbR repressor modulated by the effector substance S-adenosylhomocysteine controls directly the transcription of a regulon involved in sulphur metabolism of *Corynebacterium glutamicum* ATCC 13032. *Mol. Microbiol*., **56**, 871–887.

Rodionov D.A., Vitreschak A.G., Mironov A.A., Gelfand M.S. (2004) Comparative genomics of the methionine metabolism in Gram-positive bacteria: a variety of regulatory systems. *Nucl. Acids Res*., **32**, 3340–3353.

Schell M.A. (1993) Molecular biology of the LysR family of transcriptional regulators. *Annu. Rev. Microbiol.*, **47**, 597–626.

Shelver D., Rajagopal L., Harris T.O., Rubens C.E. (2003) MtaR, a regulator of methionine transport, is critical for survival of group B streptococcus in vivo. *J. Bacteriol*., **185**, 6592–6599.

Sperandio B., Polard P., Ehrlich D.S., Renault P., Guedon E. (2005) Sulfur amino acid metabolism and its control in *Lactococcus lactis* IL1403. *J. Bacteriol*., **187**, 3762–3778.

Thompson J.D., Gibson T.J., Plewniak F., Jeanmougin F., Higgins D.G. (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucl. Acids Res*., **25**, 4876–4882.

PHYLOGENETIC ANALYSIS OF COG1649, A NEW FAMILY OF PREDICTED GLYCOSYL HYDROLASES

***Kuznetsova A.Y.*[*1], *Naumoff D.G.*[2]**
[1] Moscow State University, Department of Bioengineering and Bioinformatics, Moscow, Russia;
[2] State Institute for Genetics and Selection of Industrial Microorganisms, Moscow, Russia
[*] Corresponding author: e-mail: anastaeae@gmail.com

Key words: COG1649, enzyme classification, protein family, α-galactosidase superfamily, protein phylogeny, DUF187, glycoside hydrolases

SUMMARY

Motivation: Recent progress in genome sequencing has resulted in accumulation of a huge number of new protein sequences. The majority of them were not and will never be characterized enzymatically. Grouping these proteins into families allows to predict their enzymatic activity or another biologic function on the basis of known homologs. If there is no enzymatically characterized relatives, then it is relevant to perform experimental study of most divergent representatives. A preliminary annotation is possible on the basis of detailed analysis of evolutionary relationship with proteins from other families. COG1649 (Pfam DUF187) is among families with no characterized members.

Results: Analyzing COG1649 we identified a new family of hypothetical glycoside hydrolases by using sequence similarity. Four subfamilies were distinguished in it on the basis of pairwise sequence comparison and phylogenetic analysis. Iterative sequence analysis revealed the relationship of the COG1649 family with the GH27, GH31, and GH36 families of glycosidases, which belong to the α-galactosidase superfamily, as well as a more distant relationship with some other glycosidase families (GH13 and GH20).

INTRODUCTION

Comparison of proteins encoded in complete genomes from major phylogenetic lineages and elucidation of consistent patterns of sequence similarities allowed the delineation of clusters of orthologous groups (COGs). Each COG consists of individual orthologous proteins or orthologous sets of paralogs from at least three lineages. Orthologs typically have the same function, allowing transfer of functional information from one member to an entire COG. Each COG is assumed to have evolved from an individual ancestral gene through a series of speciation and duplication events (Tatusov *et al*., 1997). The COG collection currently consists of 138,458 proteins, which form 4873 COGs and comprise 75 % of the 185,505 (predicted) proteins encoded in 66 genomes of unicellular organisms (Tatusov *et al*., 2003).

Glycoside hydrolases or glycosidases (EC 3.2.1.-) are a widespread group of enzymes, hydrolyzing the glycosidic bonds between two carbohydrates or between a carbohydrate and an aglycone moiety. Currently, about 30,000 sequences of glycosidases and their homologues are known. Some related families of glycosidases, having the same molecular mechanism of hydrolyzing reaction, have been combined into clans. Glycosidases catalyze hydrolysis of the glycosidic bond of their substrates via two general

mechanisms, leading to either inversion or overall retention of the anomeric configuration at the cleavage point.

International classification of glycoside hydrolases groups all known enzymes into more than 100 families on the basis of sequence similarity (http://www.cazy.org/CAZY/). Proteins of the same family share the same catalytic residues, mechanism and fold. The enzymatic polyspecificity of many glycosidase families makes it reasonable to divide them into subfamilies. In several cases proteins from different families, which do not belong to same clan, appeared to be evolutionary relative to each other. For example, GH27, GH31, and GH36 have sequence homology, similar composition of the active center, common catalytic mechanism, and a common $(\beta/\alpha)_8$-barrel type tertiary structure of the catalytic domain. That is why it was suggested to group them into α-galactosidase superfamily. Several subfamilies were distinguished in each family based on pairwise sequence comparison and phylogenetic analysis. Proteins having identity not less than 30 % are considered to belong to the same subfamily. In contrast to GH27 and GH31 families, GH36 family turned to be a polyphyletic group (Naumoff, 2004). The enzymes of this superfamily have distant relationship with proteins of other families including GH13 and GH97 families of glycoside hydrolases and COG1649 (Naumoff, 2004; Naumoff, 2005). COG1649 is a family of enzymatically uncharacterized proteins encoded by ORFs from several bacterial genomic projects. A glycosidase activity at least for some members of COG1649 can be suggested.

In the present communication, by phylogenetic analysis of COG1649 family we distinguished four subfamilies, and iterative database search allowed us to reveal interfamily relationship.

METHODS

The COG1649 family members sequences were retrieved from GenPept database using PSI-BLAST program (position-specific iterative BLAST, which combines BLAST search with profile analysis). *E*-value equaled 0.001 on each search. Protein family analysis was performed using standard methods (Naumoff, 2006). Particularly, the phylogenetic trees were built using the Neighbor-Joining and Maximum Parsimony algorithms (PHYLIP package).

RESULTS AND DISCUSSION

Sequences of all seven proteins, which belong to the COG1649 family (http://www.ncbi.nlm.nih.gov/COG), were used for PSI-BLAST screening of the GenPept database. For example, when *Crocosphaera watsonii* protein CwatDRAFT_6023 (GenPept accession number EAM52769.1) was the query sequence, the first iteration retrieved 103 sequences. The second iteration found 5 new COG1649 proteins. Several divergent representatives of the family were used for additional PSI-BLAST searches. The obtained set of proteins was used for comparative analysis.

An analysis of the order of the display sequence during searches by PSI-BLAST was used for a preliminary division of a family into subfamilies. The latter was defined as a group of proteins that are displayed at the top of the list in a PSI-BLAST query results. Depending on particular criteria of the protein similarity used, the algorithm can split a family into a larger or smaller number of groups of proteins. Allowing the members of a particular subfamily to have not less than 30 % sequence identity, we distinguished 4 subfamilies.

Phylogenetic analysis was used in order to verify the obtained subfamilies and to clarify their boundaries. The monophyletic status was used as a criterion for the final definition of a subfamily. We checked the clustering of the family members in the phylogenetic tree (Fig. 1.). The Maximum Parsimony and the Neighbor-Joining trees

show similarities in the tree topology, suggesting the correct interpretation of the evolutionary events. The family can really be divided in 4 subfamilies of a different size. Considering any subfamily as an outgroup, Neighbor-Joining tree showed that other subfamilies form monophyletic groups. The least bootstrap value for Neighbor-Joining tree equals 77 % and is relatively high. It is important that there is 1 pair of subfamilies (the first and the fourth subfamilies) that make one cluster on the Maximum Parsimony tree with significant bootstrap support. This suggests that the first and the fourth subfamily members are closer to each other than other subfamilies and they have less evolutionary distance.

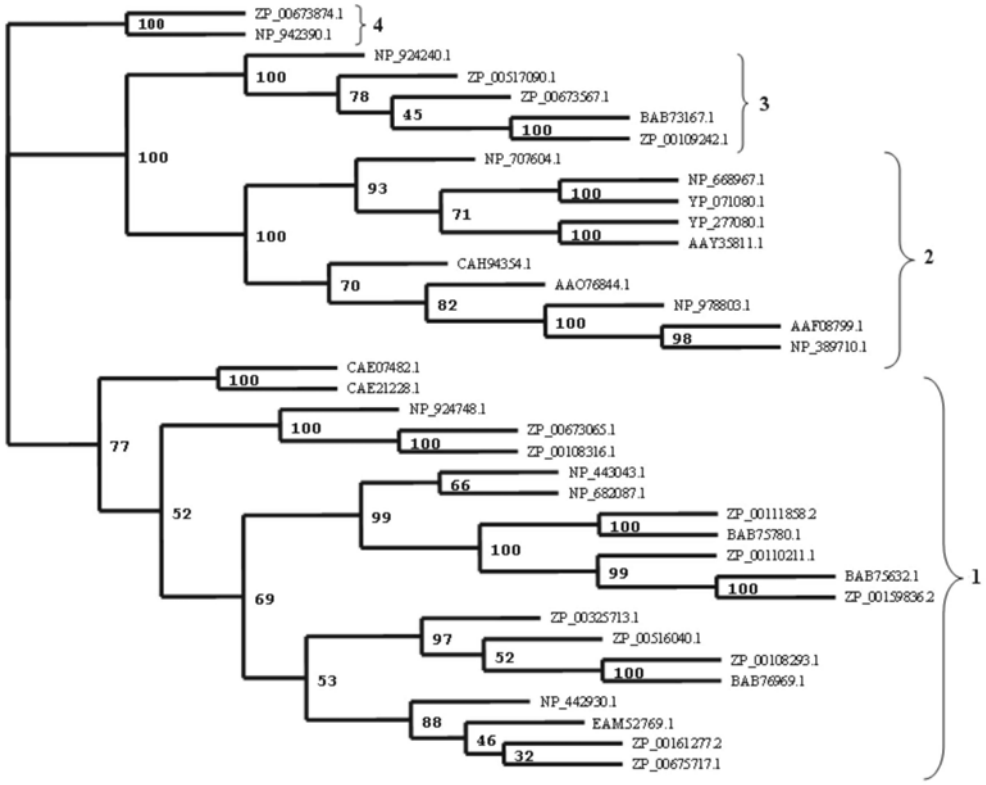


Figure 1. The COG1649 phylogenetic tree (Neighbor-Joining algorithm), visualized by TreeView program. The IDs shown on the tree are GenPept numbers. Subfamilies are indicated in the most right.

Analyzing the multiple alignment it can be seen that there are forty-seven 90 %-conservative positions in the alignment. Each subfamily has its diagnostic conservative positions.

Analysis of second iteration of PSI-BLAST showed following results: each subfamily members reveal some members of GH-families, which belong to α-galactosidase superfamily. Members of first family yielded members of GH13, GH31, GH36 families, members of second subfamily - members of GH13, GH36, GH66 families, members of the third and fourth subfamilies yielded members of GH36 family.

Analysis of the multiple alignment (Fig. 2) allowed us to find an aspartic acid amino acid residue, which is considered to be a nucleophile in catalytic center of α-galactosidase superfamily members, so we can predict that this residue is a nucleophile in the catalytic center of COG1649 members too. Our results suggest that at least some proteins of COG1649 can have a glycosidase activity. We can predict for them a $(\beta/\alpha)_8$-barrel fold of the catalytic domain and retaining mechanism of the glycoside bond hydrolysis.

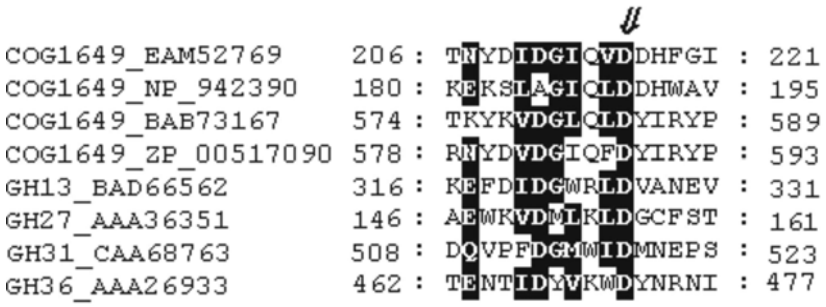


Figure 2. Fragment of a multiple sequence alignment of several representatives of COG1649 and their homologues from other families. The arrow indicates a conserved Asp residue. It plays a role of nucleophile in the active center, as it was shown for glycosidases from several families. Numbers before and after the fragments of sequences show their place in the protein sequence.

The current work was the first to be dedicated to COG1649. It allowed to show its subfamily structure and interfamily relationship; to predict 3D structure of catalytic domain and a glycosidase activity. In order to verify the enzymatic activity we suggest studying properties of several COG1649 members. It is reasonable to use one representative from each of four found subfamilies.

REFERENCES

Naumoff D.G. (2004) The α-galactosidase superfamily: sequence based classification of α-galactosidases and related glycosidases. *Proceedings of The Fourth International Conference on Bioinformatics of Genome Regulation and Structure. July 25-30, 2004.* Novosibirsk. Russia, **1**, 315–318.

Naumoff D.G. (2005) GH97 is a new family of glycoside hydrolases, which is related to the α-galactosidase superfamily. *BMC Genomics*, **6**, Art.112.

Naumoff D.G. (2006) Phylogenetic analysis of a protein family. *Zbio*, **1**, Art.3 (http://zbio.net/bio/001/003.html).

Tatusov R.L., Koonin E.V., Lipman D.J. (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.

Tatusov R.L., Fedorova N.D., Jackson J.D., Jacobs A.R., Kiryutin B., Koonin E.V., Krylov D.M., Mazumder R., Mekhedov S.L., Nikolskaya A.N., Rao B.S., Smirnov S., Sverdlov A.V., Vasudevan S., Wolf Y.I, Yin J.J, Natale D.A. (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, Art.41.

EVOLUTIONARY CONSTRUCTOR: A PACKAGE FOR MODELING COEVOLUTION OF UNICELLULAR ORGANISMS

Lashin S.A.*[*], *Likhoshvai V.A., Kolchanov N.A., Matushkin Yu.G.
Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia
[*] Corresponding author: e-mail: lashin@bionet.nsc.ru

Key words: evolution, population, mathematical model, computer analysis

SUMMARY

Motivation: There are models studying interactions within and between populations as a cause of evolutionary changes and speciation. Numerous methods and approaches have been developed for modeling such processes. Their drawbacks are that continuous approaches neglect polymorphism, and methods involving population portraits can consider but very small populations.

Results: We developed a method for modeling evolution in populations of unicellular organisms forming a trophic web and a program package for implementing this method. The software has a graphic user interface. It allows modeling the behavior of associations of polymorphic populations, mutations, horizontal transfer, and environmental changes. It can be applied to evolutionary studies.

INTRODUCTION

Computer modeling of evolutionary processes ranks among the main branches of mathematical modeling in biology. Evolution-modeling methods can be divided into portrait modeling methods (individual-oriented) and generalized modeling (Levchenko, 1993; Menshutkin, 2003). Although portrait models provide detailed description of complex processes, they are difficult for computing, because each individual should have a personal object (equation etc.) for its description. Most of generalized modeling methods are continuous, although there are statistical ones. They are simpler for computing but more rigid and less detailed. Currently existing methods allow modeling processes at the levels from genotype to population (Fog, 2000). Nevertheless, there are processes beyond the scope of any method known to us. They include horizontal transfer. Probably, this is related to the fact that most methods provide static models, whose structure remains intact in the course of computing, whereas horizontal transfer implies emergence of new populations. Besides, most methods deal with a single facet of evolutionary modeling rather than with the whole set of evolutionary processes (Vincent, 1996).

We propose a novel method of evolutionary modeling and the "Evolutionary constructor" package, implementing this method. The proposed method of stepwise mimicking modeling handles both generalized and portrait models, allowing shift from "generalization" to "portraitness" and *vice versa*.

METHODS AND ALGORITHMS

The proposed method allows modeling coexistence of populations of unicellular organisms in an association within a limited volume (environment). Each of them consumes certain substrates and excretes waste products, which can be consumed by other populations. Also, there is supply of common nonspecific substrates consumed by several or all populations. Some substrates can inhibit cells of certain populations. The effect of a substrate is determined by the presence of a certain "gene" or, more precisely, the value of the corresponding trait. A cell is assigned to a certain population just according to the set of consumed/excreted substrates. The populations form trophic webs of various structures. There may be both intraspecific and interspecific competition (In this model, populations are interpreted as species). Cells can experience mutations concerning substrate consumption and production. Such mutations may affect the sizes of the populations and modes of their existence. Horizontal transfer of genetic information between cells can occur. To implement these functions, the following objects and processes (interactions between objects) are distinguished:

Objects

A population is a set of genetically uniform haploid individuals. The genes of an individual are divided into two groups. Genes of group 1 determine quantitative traits, such as the rate of substrate consumption by an individual (gene–substrate correlation). Genes of group 2 determine such quantitative traits as the rate of production (gene–product correlation). The genotype of an individual determines its need for substrates, ability to compete for substrates, and use substrates for propagation and production. To cope with the huge volume of computing, characteristic of portrait models, we introduce the term "genetic spectrum of a population". It denotes the distribution of the occurrence of a gene in the population (Fig. 1). The definition above is used for formalization of species (population, strain etc) concept. It is well known that there are some definitions of species even for bisexual organisms. To define the species (population etc.) for unicellular ones is much more difficult in view of more traits variability. Concordance (in some limits) of genetic spectrums leads to belonging of organisms to a one species (population etc.). For more formal definition of these limits we also introduce the notion of the threshold value of a trait in the spectrum. This means that a trait is considered to be absent from a cell if its value is below the threshold. For example, in Fig. 1, Gene1 is considered to absent from 10 % of the population. Cells are assigned to one population if their sets of present and absent (with regard to threshold values) traits match. A change of the spectrum of a population is considered a mutation (Fig. 2). If the spectrum partially crosses the threshold, part of the cells can come to another population (segregation of the population; Fig. 3). In this case, if a population with a similar spectrum existed before the segregation, the subpopulation joins it; otherwise, a new population arises. So, this methodology allows modeling of unicellular populations' evolution.

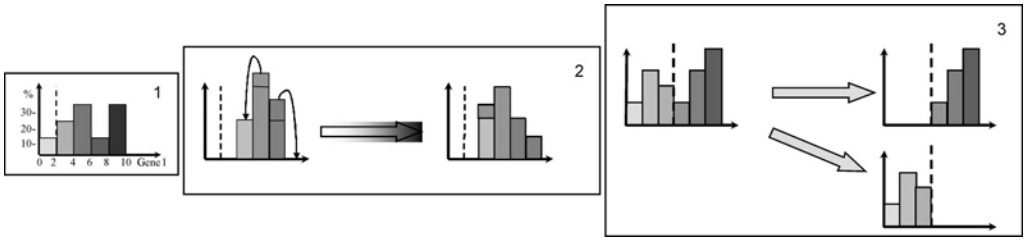


Figure 1. The environment is a closed volume containing substrates, products, and populations – 1, 2, 3.

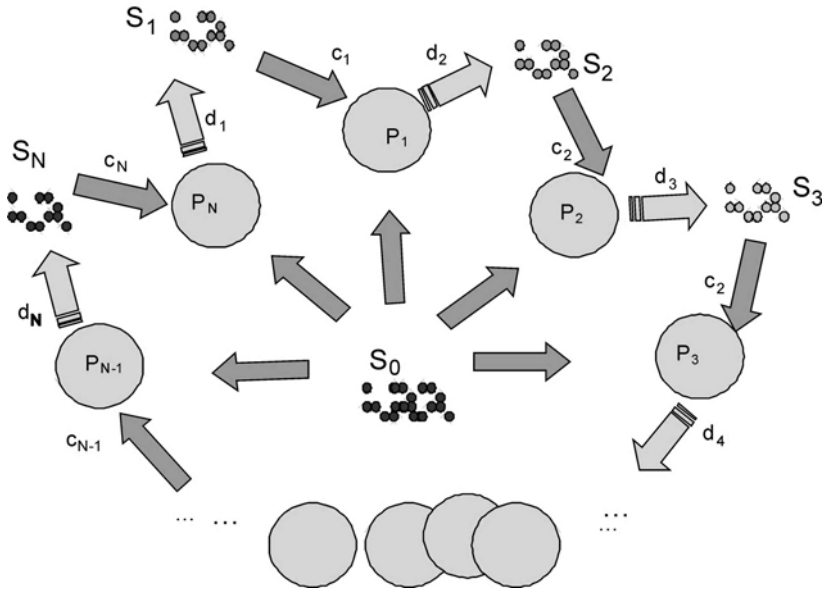


Figure 2. Interaction between populations and the environment by the example of a trophic ring (a special case of trophic web). Substrates are designated as S_i (a nonspecific substrate is designated as S_0); c_i is the rate of consumption of the corresponding substrate; d_i is the rate of production of the corresponding substrate.

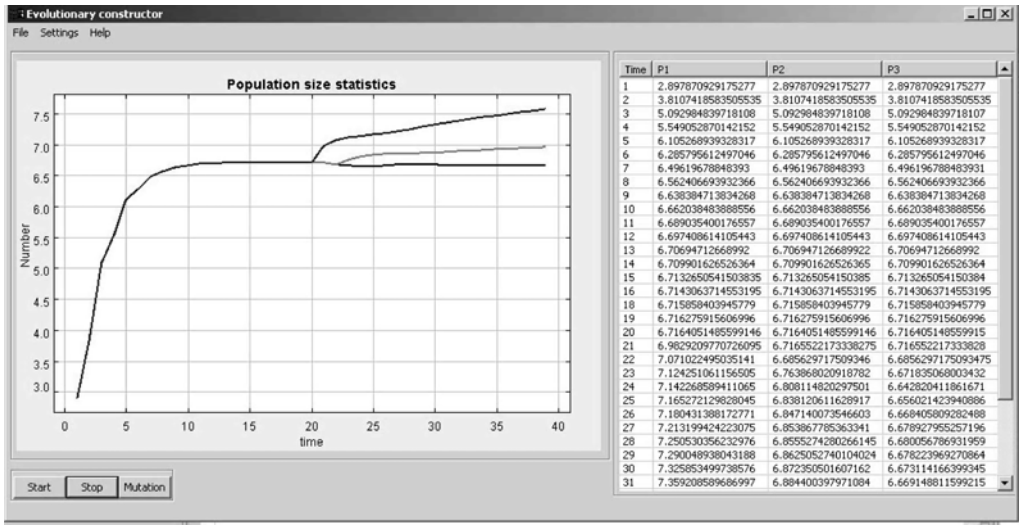


Figure 3. "Evolutionary constructor" window.

Interaction between objects

Substrate consumption by populations. Individuals of each population consume the required substrates from the environment. The amount of consumed substrate depends on the demand of the individual and on the availability of the substrate in the environment. With an excess of the substrate, the demands of all individuals are met; otherwise, individuals compete for the substrate. The algorithm allows adjustment of demands.

Population size change. Reproduction rate of a population depends on the amount of the substrate consumed. With its shortage, it is close to zero. Population size also depends

on mortality. It can be accompanied by a change of the genetic spectrum of the population. Such changes are related to evolutionary advantages of individuals more efficiently assimilating substrates. To calculate the growth of a population, we divide it into a set of subpopulations. Individuals of each of the subpopulations are genetically uniform; i.e., subpopulations are monomorphic. The following equations can be used for calculating the growth of a monomorphic population:

$$F_1(\vec{S},\vec{C},P) = \sqrt{c_0 s_0(P) \cdot \sum_{i=1}^{N} c_i s_i(P)} - k_{death} \cdot P^2, \quad (1)$$

where S is the concentration vector of the substrate **consumed** by the population, proportional to the subpopulation size; C, vector of traits determining consumption rate, calculated for each subpopulation from the spectrum; P, subpopulation size; and k_{death} is the mortality index of the population. This formula takes into account the necessity of common substrate to a population growth.

$$F_2(\vec{S},\vec{C},P) = a_{basal} \cdot P - \sqrt{\sum_{i=1}^{N} c_i s_i(P)} - k_{death} \cdot P^2, \quad (2)$$

where a_{basal} is the natural growth of the population. Other designations follow (1). This formula describes the inhibitory influence of substrates.

$$F_2(\vec{S},\vec{C},P) = \frac{a_{basal}}{1 + \left(\sum_{i=1}^{N} c_i s_i(P) \right)^{\gamma}} - k_{death} \cdot P, \quad (3)$$

where γ is the nonlinearity degree of the effect of inhibiting substrates on population growth. Other designations follow (2). This formula describes the complex influence of substrate (small concentrations result in positive growth of population size, large concentrations inhibit this growth).

Flux in the environment. The environment is characterized by supply of the nonspecific substrate from the outer space and discharge to the outer space.

A mutation is described as a change of the genetic spectrum of a population. Horizontal transfer is described as an emergence or disappearance of a trait.

IMPLEMENTATION AND RESULTS

The described modeling method was implemented in the "Evolutionary constructor" program package. This software permits the user to construct and compute models, handle mutations and horizontal transfer, vary environmental parameters, and change growth equations. Statistics is presented in graphic and text forms. Models can be saved in the internal format. Statistics can also be saved separately in the text format.

DISCUSSION

The software for simulating evolution developed by us on the grounds of a pioneering method can be applied to a wide range of evolutionary problems. Its main advantages are:

possibility of development and dynamic variation of various parameters of models and the possibility to change the degree of "portraitness" of a model by using population genetic spectra. "Evolutionary constructor" was used for simulation of several evolutionary events, including horizontal transfer (Lashin *et al*., 2006).

ACKNOWLEDGEMENTS

The work was supported by the Russian Foundation for Basic Research (Nos 04-01-00458, 05-04-49068, 05-07-90274, 06-04-49556), Project "Evolution of molecular-genetic systems: computer analysis and modeling" of the RAS Presidium program "Biosphere origin and evolution" No. 10104-34/P-18/155-270/1105-06-001/28/2006-1. Project "Computer modeling and experimental constructing of gene networks" of the RAS Presidium program of molecular and cell biology.

REFERENCES

Fog A. (2000) Simulation of Evolution in Structured Populations: The Open Source Software Package Altruist. Biotech Software & Internet Report. **1**(5), 226–229.

Lashin S.A., Likhoshvai V.A., Kolchanov N.A., Matushkin Yu.G. (2006) Modeling of horizontal gene transfer in prokaryotic populations with the "Evolutionary Constructor" program package. *Thiss issue.*

Levchenko V.F. (1993) Models in biology evolution theory. *Spb, Nauka*, 384.

Menshutkin V.V. (2003) Computer simulation of different types of evolutionary process. *Zh. Obshch. Biol*., **64**(4), 328–336.

Vincent T.L. (1996) Modeling and management of evolutionary component in biological systems. *Ecol. Model*., **92**(1), 145–153.

MODELING OF HORIZONTAL GENE TRANSFER IN PROKARYOTIC POPULATIONS WITH THE "EVOLUTIONARY CONSTRUCTOR" PROGRAM PACKAGE

Lashin S.A.*[*], *Likhoshvai V.A.*, *Kolchanov N.A.*, *Matushkin Yu.G.
Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia
[*] Corresponding author: e-mail: lashin@bionet.nsc.ru

Key words: evolution, population, mathematical model, computer analysis, horizontal transfer, mutations

SUMMARY

Motivation: By now, many models have been developed for theoretical study of evolution in populations, mutations, and intraspecific and interspecific struggle (Fog, 2000). Nevertheless, horizontal gene transfer, being a biological phenomenon of great evolutionary importance, is poorly reflected in mathematical models.

Results: We consider a model of horizontal gene transfer in populations of unicellular organisms forming a closed trophic chain. The model has been analyzed under both steady-state and variable environmental conditions. Possibility of adaptive evolution by horizontal gene transfer has been confirmed. The simulation demonstrates an intriguing phenomenon: extinction of part of the trophic ring with survival of a "parasite" population in the remaining subring.

INTRODUCTION

Computer modeling of evolutionary processes ranks among the main branches of mathematical modeling in biology. Evolution modeling methods can be divided into individual-oriented portrait modeling methods (Dieckmann *et al.*, 2004) and generalized modeling (e.g., prey-predator model). Currently existing methods allow modeling processes at the levels from genotype to population. Nevertheless, we do not know any reports on modeling horizontal transfer. Probably, this is related to the fact that most methods provide static models, whose structure remains intact in the course of computing, whereas horizontal transfer implies emergence of new populations. Here we use a pioneering approach implemented in the "Evolutionary constructor" program package for simulating horizontal transfer in populations of unicellular organisms.

METHODS AND ALGORITHMS

The model is constructed according to the method described in (Lashin *et al.*, 2006). It is applicable to modeling an association of populations, whose interaction with one another is mediated by trophic chains. The populations are confined to a limited volume (environment). Each of them consumes certain substrates and excretes metabolites, which can be consumed by other populations. There is a supply of a common nonspecific

substrate consumed by all populations. For modeling populations, we introduce the term "genetic spectrum of a population", which is the vector of distribution of the occurrence of a gene in the population (for details see Lashin *et al*., 2006). This spectrum allows treatment of a population as a single object, as in generalized modeling methods. Calculation of population growth involves special algorithms for properly modeling intra- and interspecific competition. Competition occurs at the level of struggle for substrates. Thus, the approach combines advantages of generalized and individual-oriented modeling methods. It allows modeling mutations, altering the rates of substrate consumption and production; environmental changes; and horizontal transfer.

Models are constructed and calculated according to this approach with the use of "Evolutionary constructor", described in (Lashin *et al*., 2006). Population growth is calculated as follows:

$$F_1(\vec{S},\vec{C},P) = \sqrt{c_0 s_0(P)\cdot\sum_{i=1}^{N} c_i s_i(P)} - k_{death}\cdot P^2, \quad (1)$$

where S is the concentration vector of the substrate **consumed** by the population, proportional to the population size; C, vector of traits determining consumption rate, calculated for each population from the spectrum; and P, population size.

IMPLEMENTATION AND RESULTS

For modeling horizontal transfer, we specified a trophic ring of seven populations. In this ring, ith population consumes the ith specific substrate and secretes ((i+1) mod 7) specific substrate No. (i+1) mod 7. In the figure, the substrates are lettered from A to G. In addition, all populations consume the common nonspecific substrate. All populations have equal rates of consumption of the nonspecific substrate and corresponding specific ones. At time zero, all the seven populations are in equilibrium. At the 10^{th} time step, horizontal transfer of a gene occurs in one cell of population 1, and it gains the capacity to consume substrate B in addition to the main substrate G. This cell founds a population 8, different from others, which rapidly grows to a permanently large size owing to its ability to consume two specific substrates. An even larger size of population 2 is determined by the fact that substrate A is produced by two populations, P1 and P8.

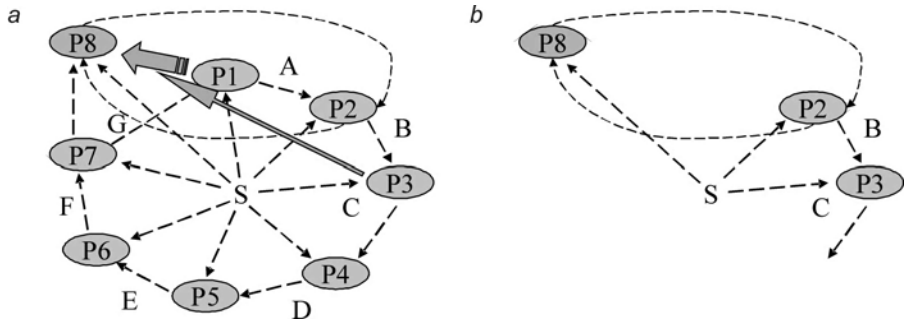


Figure 1. Schematic representation of the trophic ring, horizontal transfer, and emergence of a new population 8 (Dashed lines correspond to substrate-product relationships; gradient lines show the horizontal transfer of gene from P3 species to P1 one and the further genesis of P8) (*a*). Schematic representation of the trophic subring after a short period of "hunger-strike" (*b*).

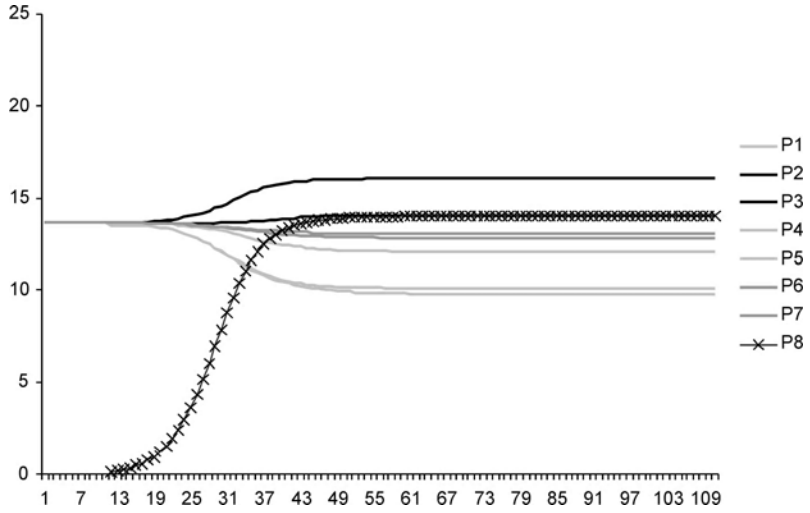


Figure 2. Development of the new population 8.

Additional factors, biologically conceivable, can yield even more dramatic results. Consider a new population in the trophic ring, appearing in the same way. After a certain time, when all populations reach the steady state, the supply of the common substrate decreases 200-fold, and then, after another period of time, the supply returns to the initial level. This may mimic alternation of rainy and dry seasons. We notice that during "starvation" some populations became irreversibly extinct (were not restored after resumption of substrate supply). Only populations participating in the trophic subring (P8, P2, P3), formed by horizontal transfer, survived. It can be explained by the fact that these populations had the most size prior the "hunger-strike", which is determined by the various factors: P8 could consume two types of the specific substrate, B and G; P2 got specific substrate from two populations – P8 and P1; P3 got the largest portion of specific substrate from P2. During starvation populations began to extinct: at first the populations P1 and P4 being in the less size have perished, then the order of dying determined by the size of population in stationary state.

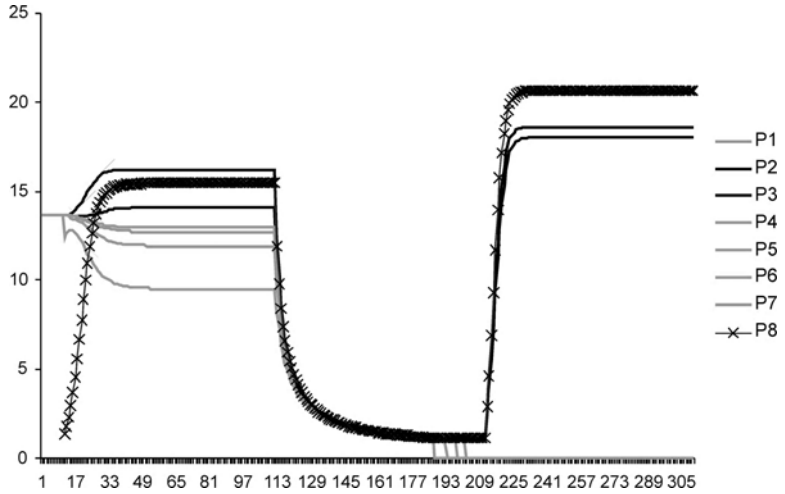


Figure 3. Variation of population sizes in the ring after emergence of a new population and subsequent reduction of common substrate supply.

DISCUSSION

Numerical modeling of horizontal transfer in unicellular organisms confirms the hypothesis that horizontal gene transfer may offer evolutionary advantage. Also, we found an intriguing phenomenon: After the death of part of the trophic ring, the remaining subring contains a population (P3) unhelpful to other members of the ring, because the substrate produced by the population is not consumed by anyone. Thus, this population is actually a parasite. As the problem is stated by us, horizontal transfer by itself cannot cause serious consequences. One population acquires better living conditions, and its size increases. However, abrupt environmental changes (common substrate concentration, supply rate, etc.) can make such exchange of genes important for positive natural selection. The survivor subring contains a "parasite" population, which may function as a variability resource for further evolution.

ACKNOWLEDGEMENTS

The work was supported by the Russian Foundation for Basic Research (Nos 04-01-00458, 05-04-49068, 05-07-90274, 06-04-49556), Project "Evolution of molecular-genetic systems: computer analysis and modeling" of the RAS Presidium program "Biosphere origin and evolution" No. 10104-34/P-18/155-270/1105-06-001/28/2006-1. Project "Computer modeling and experimental constructing of gene networks" of the RAS Presidium program of molecular and cell biology.

REFERENCES

Dieckmann U., Doebeli M., Johan A.J., Metz, Tautz D. (2004) Adaptive Speciation. Cambridge University Press, 446 p.

Fog A. (2000) Simulation of Evolution in Structured Populations: The Open Source Software Package Altruist. Biotech Software & Internet Report. **1**(5), 226–229.

Lashin S.A., Likhoshvai V.A., Kolchanov N.A., Matushkin Yu.G. (2006) Evolutionary Constructor: a package for modeling coevolution of unicellular organisms. *This issue.*

Menshutkin V.V. (2003) Computer simulation of different types of evolutionary process. *Zh. Obshch. Biol.*, **64**(4), 328–336.

MOLECULES *VERSUS* MORPHOLOGY IN OLIGOCHAETA SYSTEMATICS

Liventseva V.**, *Kaygorodova I.
Limnological Institute, SB RAS, Irkutsk, Russia
* Corresponding author: e-mail: live@lin.irk.ru

Key words: oligochaeta, species flocks, morphology based phylogeny

SUMMARY

Motivation: With accumulation of new data a revision of existing classification of oligochaete worms is indispensable, in view of construction of a phylogenetically justified structure of taxa.

Results: The analysis of a complex significant morphological characters has allowed to find out a degree of change of characters during evolution and this has allowed to define a direction of evolutionary transformations and related connections within family and higher levels of taxa.

INTRODUCTION

Family Lumbriculidae is among the most diverse species groups of Lake Baikal. World lumbriculid fauna accounts for ca. 120 species, almost half of it are described from Baikal. Recent studies on evolution of Baikalian oligochaete based on molecular data mainly (Kaygorodova, 2000) were find out some monophyletic groups of closely related species of family Lumbriculidae inhabiting the lake. Species flocks are the most attractive objects for phylogenetic study, because the rate and mode of speciation may provide the clues to the direction of evolutionary processes and to mechanisms of generation of such a vast biodiversity. Given the construction of systematics classification still is not possible without attraction of the morphological data (Wiens, 2004) the present study attempts to apply modern phylogenetic methods on morphological data to resolve difficult points of oligochaete systematics. Comparison of morphology and molecular based phylogenies of Lamprodrilus and Rhynchelmis species flocks of Baikalian Lumbriculidae were done.

METHODS

The similar set of species as in previous molecular study of I. Kaygorodova consists of 26 species belonging to all 6 genera of Baikalian Lumbriculidae. Set of all world known species of genus *Rhynchelmis* was analyzed separately. Morphological analysis for Baikalian and non-Baikalian species was done on real biological material and literary descriptions respectively. Morphology based phylogeny was inferred with parsimony analysis using PAUP ver. 4.0 (Swofford, 1998).

RESULTS AND DISCUSSION

The accurate analysis of morphological data has allowed picking out the sequences by 46 discrete morphological (phenetic and anatomic) characters for each taxon.

The phylogenetic analysis of morphological data set resulted in a single most parsimonious tree, (right tree on Fig. 1). The strict consensus tree is fully resolved. Comparison of molecular and morphology based phylogenies (Fig. 1) revealed following:

(a) Species of three genera *Styloscolex*, *Stylodrilus* and *Lamprodrilus* form monophyletic groups.

(b) Species belonging to two independent genera (*Lamprodrilus* and *Teleuscolex*) clustered always together in both molecular and morphology based trees.

(c) Unique Baikalian species *Agriodrilus vermivorus* (the only predatory one among World known fauna of Lumbriculidae) has controversial position on different phylogenies. On the one hand, molecular tree proves previous morphology based hypotheses (Brinkhurst, 1989; Cook, 1971) placing it in Lamprodrilus group. On the other hand, this species clustered to Styloscolex species group on morphology based phylogeny.

(d) Baikalian representatives of genus *Rhynchelmis* are diverged on two different clades.

Polyphyletic nature of the genus supported by results of additional morphology based phylogenetic analysis inside of separately taken all presently known *Rhynchelmis* species (Fig. 2).

(e) Morphology based phylogeny proved that morphological characters of all lumbriculid genera evolved in one direction. It could be evidence of adaptation to a common inhabitancy.

Analysis of *Rhynchelmis* data allowed clearly recognizing the set of most informative morphological characters. According to high value of index of compatibility (CI) these characters are length of setae, presence of penial seta, presence of additional formations at a man's sexual aperture (penial bulbs, "saddle"), direction of sperm funnels, size of sperm funnels, amount of spermathecae, connection of spermatheca with intestine. It is recommended that these attributes will be taken into account at systematics and for construction of classification keys of genus *Rhynchelmis*.

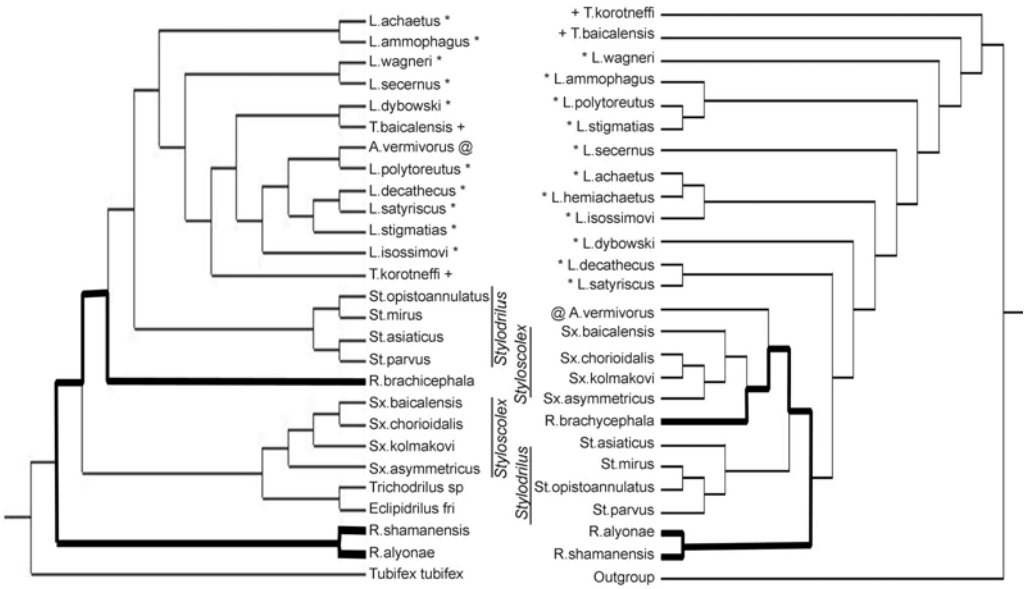


Figure 1. Intergeneric relations between baikalian Lumbriculidae proposed in a phylogenetic analysis of molecular data (left) and morphological data (right). Representatives of different genera belonging to "Lamprodrilus" group marked as following: *Teleuscolex* – "+", *Agriodrilus* – "@", *Lamprodrilus* – "*". Lineages of *Rhynchelmis* species set off in bold.

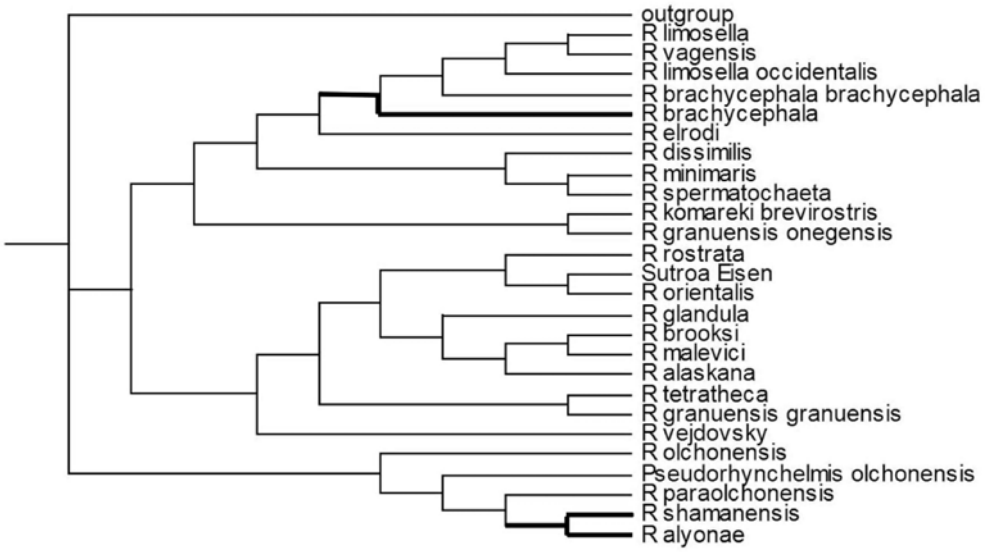


Figure 2. Scheme of morphology based phylogeny of genus *Rhynchelmis*, strict consensus of optimal trees.

REFERENCES

Brinkhurst R.O. (1989) A phylogenetic analysis of the Lumbriculidae (Annelida, Oligochaeta). *Can. J. Zool*., **67**, 2731–2739.

Cook D.G. (1971) Family Lumbriculidae. In Brinkhurst R.O, Jameison B.G.M. (eds), *Aquatic Oligochaeta of the World*. Oliver and Boyd, Edinburgh. pp. 201–285.

Kaygorodova I.A. (2000) Molecular phylogenetic study of evolutionary history of Baikalian Lumbriculids (Oligochaeta, Annelida). Review of Doctoral Thesis. Novosibirsk.

Swofford D. (1998) Paup 4.0 beta version, Laboratory of Molecular Systematics Smithsonian Institution, Massachusetts.

Wiens J.J. (2004) The role of morphological data in phylogeny reconstruction. *Syst. Biol*., **53**(4), 653–661.

NEW FAMILY OF LTR RETOTRANSPOSABLE ELEMENTS FROM FUNGI

***Novikova O.*[*1], *Fursov M.*[2], *Blinov A.*[1]**
[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia; [2] Novosibirsk Center of Information Technologies "UniPro", Novosibirsk, 633159, Russia
[*] Corresponding author: e-mail: novikova@bionet.nsc.ru

Key words: fungi, mobile elements, LTR retrotransposons, distribution, evolution, computer analysis

SUMMARY

Motivation: The transposable elements most commonly isolated from fungi to date are retrotransposons. Fungal LTR retrotransposons showed the high variability and presence of great number of different groups and families. These elements resemble the vertebrate retroviruses in their genomic organization and replication cycle. Investigated fungi revealed considerable differences in the structure of retrotransposable elements population in genomes even for closely related species. The differences suggest that the factors shaping retrotransposon evolution have differed in the distant species. Use of data generated by different genome sequencing projects to identify LTR retrotransposons in distant fungal species can give insights into evolution of mobile elements and their hosts interactions.

Results: We described two Ty3/gypsy LTR elements from *Aspergillus fumigatus* and *Aspergillus nidulans* which formed a new common branch on the phylogenetic tree of LTR retrotransposons. They appear to form a new family of fungal elements. Further screening of 11 fungal genomes revealed presence of elements belonging to the same new family not only in ascomycetes but also in basidiomycetes. Altogether 6 additional elements have been found.

Availability: http://genome.unipro.ru/.

INTRODUCTION

Retrotransposons are a significant component of many eukaryote genomes. They can be subdivided into two large groups, long-terminal repeat (LTR) and non-LTR retrotransposons, based on their overall structure. Most LTR retrotransposons can be classified into one of two distinct groups, the Ty1/copia group or the Ty3/gypsy group, based on their reverse transcriptase sequences and other structural features (Xiong, Eickbush, 1990).

The phylogenetic analysis of retrotransposons was initially based on reverse transcriptase (RT) sequences and it was found 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 elements and using only this domain for tracing of LTR retrotransposons phylogeny is not sufficient. Therefore, later integrase (Int) and ribonuclease H (RNaseH) were used in phylogenetic analysis of LTR retrotransposons (Malik, Eickbush, 1999; Malik, Eickbuch, 2001).

The studies demonstrated that Ty3/gypsy group could be further subdivided into eight clades in a phylogeny based on RT, RNase H, and Int domains: Mdg3, Cer1, Athila, Mag, Osvaldo, Gypsy, Mdg1, and Ty3 clades (Malik, Eickbush, 1999).

Majority of currently known Ty3/gypsy LTR retrotransposons from fungal genomes belong to the clade named Ty3. In present study, we discovered a new Ty3/gypsy family from both ascomycetes and basidiomycetes fungi using the computer-based approach. The newly identified family belongs to the Ty3 clade like the other majority of known fungal Ty3/gypsy retroelements. However, this family is noticeably different from other fungal Ty3/gypsy elements. Phylogenetic analysis showed that the new family formed a distantly separated branch in Ty3 clade.

METHODS

Original sequences of *Afut4* from *A. fumigatus* and *Dane3* from *A. nidulans* were evaluated by UniPro Genome Browser through total genome screening and further analysis.

UniPro Genome Browser allows performing purposeful searching of LTR retrotransposons according several parameters using plug-in called "Find ME". UniPro Genome Browser is available on web-site http://genome.unipro.ru/. All plug-ins are freely available on http://genome.sourceforge.net/index.shtml. As the first step, the search of reverse transcriptase profile was carried out using HMM Search algorithm based on hidden Markov chains. Profile was built using UniPro Genome Browser HMM Search plug-in or HMMER, a free Unix command line tool which is available on web-site http://hmmer.wustl.edu/ (Durbin *et al*., 1998). The results of HMM search were saved as annotations for sequence. For every signal found in a previous step, long terminal repeats (LTRs) were searched. Only those repeats that were close to HMM signal and surround it were selected. The resulting repeats were also saved as an annotation for sequence in a separate group. Finally, several parameters were used for filtering of results from two previous steps: search for intact open reading frames inside the sequences flanked by putative LTRs; alignment and sorting according to similarity.

To screen other genomes we used tblastn search program (Altschul *et al*., 1990) from the National Center for Biotechnology Information (NCBI: www.ncbi.nlm.nih.gov) fungal genomes database. Protein sequence of *Afut4* RT domain was used as query.

Amino acid sequences of the RT domains of each element were aligned using CLUSTALX (Thompson *et al*., 1994), followed by manual gap adjustments. Phylogenetic trees were generated by Neighbor-Joining method using MEGA2 software package (Kumar *et al*., 2001). The significance of the various phylogenetic lineages was assessed by bootstrap analysis. All parameters used in both programs were default.

RESULTS AND DISCUSSION

The elements *Afut4* and *Dane3* presumably belonging to the new family of fungal Ty3/gypsy LTR retrotransposons were detected in genomes of two Aspergillus species, *Aspergillus fumigatus* (*Afut4*) and *Aspergillus nidulans* (*Dane3*), during the whole genome screening. Phylogenetic analysis based on RT amino acid sequences showed that these elements formed common branch and seem to be distant from other known fungal Ty3-like elements (Fig. 1).

Newly identified *Afut4* element carries two putative open reading frames interrupted by single stop-codon each. However, it is highly possible that *Afut4* could retrotranspose due to the presence of intact RT, Int and RNaseH domains in the protein product of the second open reading frame. *Dane3* element was found only as one open reading frame, coded a protein with RT, Int and RNaseH enzymatic domains, and lacked flanked long repeats. Nevertheless, the amino acid sequences of RTs from *Afut4* and *Dane3* were more than 76 % similar.

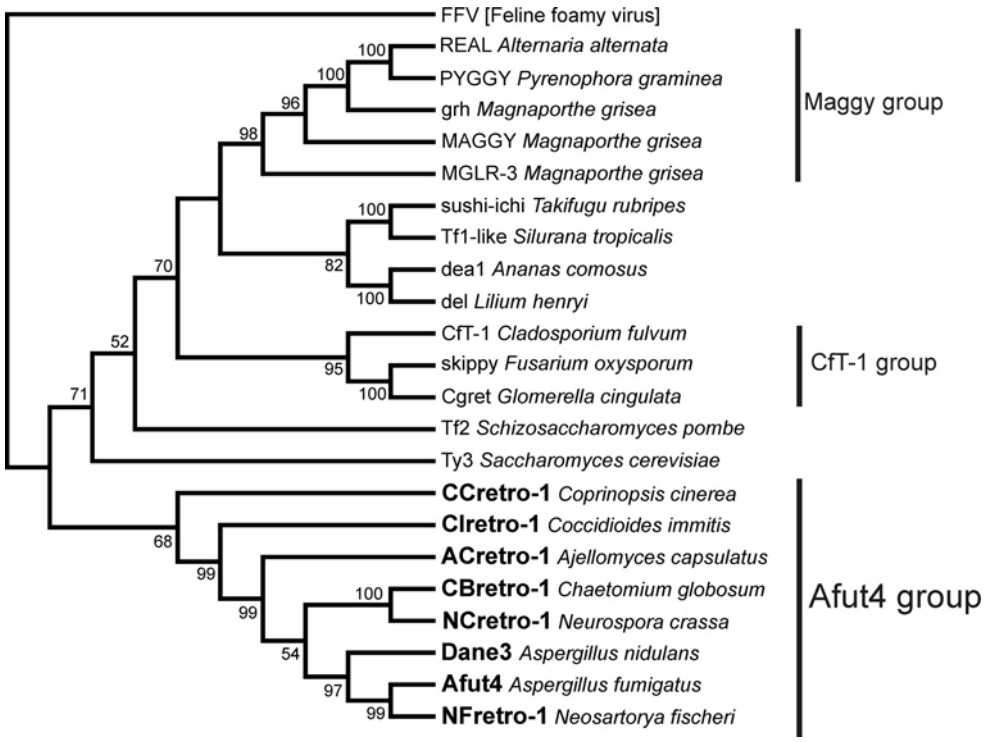


Figure 1. Phylogenetic tree of Ty3 group based on the sum of amino acids in the RT, Int and RNaseH domains (approximately 700 amino acid positions): elements found in present investigation – bold. The bootstraps with number less than 50 are not shown.

We performed tblastn search for related elements in the genomes of other fungi which are available as whole or partial genomic sequences from the National Center for Biotechnology Information.

The list of searched genomes and the results of our searches are summarized in Table 1.

In total, elements from the same family were detected in six searched genomes whereas five fungi showed no matches with query sequence. No matches were detected in search against genomic sequence of *A. terreus*, another *Aspergillus* fungus. This could be explained by an incomplete genomic sequence represented in database and a low copy number of investigated elements. Only two copies were detected in *A. fumigatus* and one in *A. nidulans*. Similarly, *Neurospora crassa* showed presence of only one degenerate copy of the element per genome.

Table 1. List of searched genomes and identified elements

Group	Species	New family
Ascomycota	*Ajellomyces capsulatus NAm1*	+ (*ACretro-1*)
	Aspergillus fumigatus Af293	+ (*Afut4*)
	Aspergillus nidulans FGSC A4	+ (*Dane3*)
	Aspergillus terreus NIH2624	no matches
	Botryotinia fuckeliana B05.10	no matches
	Chaetomium globosum CBS 148.51	+ (*CBretro-1*)
	Coccidioides immitis	+ (*CIretro-1*)
	Gibberella zeae PH-1	no matches
	Magnaporthe grisea 70-15	no matches
	Neosartorya fischeri NRRL 181	+ (*NFretro-1*)
	Neurospora crassa	+ (*NCretro-1*)
Basidiomycota	*Coprinopsis cinerea okayama7#130*	+ (*CCretro-1*)
	Phanerochaete chrysosporium	no matches

It is intriguing that both Ascomycota and Basidiomycota fungi show the presence of elements related to the *Afut4*. *Afut4*-like elements were revealed in *Coprinopsis cinerea okayama7#130*, which is representative of Basidiomycota. At the same time, the elements from this family were not found in *Phanerochaete chrysosporium* genome.

The similarity averaged 59.6 % among amino acid sequences of RT, Int and RNaseH domains from nine newly isolated elements. The similarity of RT domain alone made up more than 75 %.

All previously described fungal elements belonging to the Ty3 group of LTR retrotransposons could be subdivided into several groups. Two groups are represented by elements from filamentous fungi. Maggy and CfT groups are clearly separated on the phylogenetic tree and formed a common branch with elements from vertebrates and plants (Fig. 1). Two known elements from yeasts *Tf2* and *Ty3* constituted their own branches. The newly identified elements did not belong to any of the known groups and formed a new branch on the phylogenetic tree. We designated this group as Afut4 group of elements (Fig. 1). Afut4 group is the first fungal group of LTR retrotransposons which includes not just elements from a single fungal taxon like Maggy or CfT-1 groups but elements from several distant taxa.

ACKNOWLEDGEMENTS

This work was supported in part by the state contract 10002-251/П-25/155-270/200404-082 and Siberian Branch of the Russian Academy of Sciences (project No. 10.4).

REFERENCES

Altschul S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Durbin R. *et al.* (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge.

Kumar S. *et al.* (2001) MEGA2: molecular evolutionary genetics analysis software. *Bioinformatics*, **17**, 1244–1255.

Malik H.S., Eickbush T.H. (1999) Modular evolution of the integrase domain in the Ty3/Gypsy class of LTR retrotransposons. *J. Virol*., **73**, 5186–5190.

Malik H.S., Eickbuch T.H. (2001) Phylogenetic analysis of ribonuclease H domains suggests a late, chimeric origin of LTR retrotransposable elements and retroviruses. *Genome Res*., **11**, 1187–1197.

Thompson J.D., Higgens D.G., Gibson T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl. Acids Res.*, **22**, 4673–4680.

Xiong Y., Eickbush T.H. (1990) Origin and evolution of retroelements based upon their reverse transcriptase sequences. *The EMBO J.*, **9**, 3353–3362.

CONCERTED EVOLUTION OF PARALOGOUS Oas1 GENES IN RODENTIA AND CETARTIODACTYLA

***Perelygin A.A.*[*1], *Zharkikh A.A.*[2], *Brinton M.A.*[1]**
[1] Biology Department, Georgia State University, Atlanta, GA 30302, USA; [2] Bioinformatics Department, Myriad Genetics, Inc., Salt Lake City, UT 84108, USA
[*] Corresponding author: e-mail: aperelygin@gsu.edu

Key words: 2'-5' oligoadenylate synthetase, concerted evolution, gene conversion

SUMMARY

Motivation: The mouse 2'-5' oligoadenylate synthetase-1B (Oas1b) gene confers resistance to disease induced by flaviviruses, including West Nile virus (Perelygin *et al*., 2002). Oas1b is a member of the Oas gene family, which is located on mouse chromosome 5. The molecular evolution of paralogous mammalian 2'-5' oligoadenylate synthetase-1 genes has not been previously described.

Results: A total of fourteen new mRNA sequences of paralogous Oas1 genes were determined in mice, rats, cattle and pigs. These sequences as well as seven Oas1 gene sequences from GenBank were mapped to genomic regions and also used to build a phylogenetic tree. The majority of the eight mouse Oas1 genes clustered with their rat orthologs. However, the differences between paralogous rodent Oas1c and Oas1d genes in each species were smaller than the distances between the corresponding orthologs suggesting a concerted evolution of these genes. A new method was developed to compare the distribution of substitutions along the Oas1 nucleotide sequences in rodent or even-toed ungulate evolutionary lineages and to quantify these differences as probabilities. The distributions for pairs of sequences were then compared using the non-parametric Kolmogorov-Smirnov test and the results suggested that the homogenization of paralogous 2'-5' oligoadenylate synthetase-1 genes was due to gene conversion.

INTRODUCTION

2'-5' oligoadenylate synthetases are important components of an interferon-mediated antiviral pathway, but are also involved in other cellular processes such as apoptosis, cell growth and differentiation, gene regulation, DNA replication and RNA splicing. The sequences of most members of the mouse 2'-5' oligoadenylate synthetase (Oas) gene family and all members of the human OAS gene family have previously been reported (Justesen *et al*., 2000; Perelygin *et al*., 2002). Four OAS genes (hOAS1, hOAS2, hOAS3 and hOASL1) are located on human chromosome 12. The murine Oas family includes eight small mOas1 genes (mOas1a through mOas1h), a medium sized mOas2 gene and a large mOas3 gene, as well as two Oas-like genes, mOasl1 and mOasl2, which contain two tandemly repeated ubiquitin-like sequences at their 3'-ends. Because only few additional mammalian 2'-5' oligoadenylate synthetase genes were previously reported, it was not known whether the gene families in other mammals were more like the one in mice or the one in humans. Two rat rOas1 genes located on rat chromosome 12 were previously detected. Also, single OAS1 gene sequences from both pig and cattle were previously submitted to GenBank. In this study, we analyzed the 2'-5' oligoadenylate synthetase gene families of cattle, pig, mouse and rat and obtained evidence that supports concerted

evolution of paralogous 2'-5' oligoadenylate synthetase-1 genes within the orders Rodentia and Cetartiodactyla.

METHODS

GenBank was searched for 2'-5' oligoadenylate synthetase genes and pseudogenes using the *Blast* program (Altschul *et al*., 1997). The *Dotmap* program was used to visually identify additional exons and pseudogenes. Partial rat, cattle and pig mRNA sequences predicted by the *bl2seq* program were used to design gene-specific primers, which were then utilized to amplify full-length sequences from commercial cDNAs. Multiple alignments of the full-length sequences were performed with the *Align* program. Manual editing was used to synchronize the positions of gaps between pair-wise alignments. The phylogenetic tree was constructed using the *njtree* program as described previously (Perelygin *et al*., 2005). The confidence of each node was estimated using a bootstrap algorithm (Zharkikh, Li, 1995).

RESULTS AND DISCUSSION

Eigth new rat 2'-5' oligoadenylate synthetase genes and two pseudogenes were predicted in the NW_047376 supercontig. Predicted sequences were used to amplify six full-length rOas1 cDNAs. Maps of mouse and rat Oas1 clusters were compared (Fig. 1) to suggest the orthologous relationships between individual genes. These relationships were validated by comparison of Oas1 tissue expression patterns in mice and rats.

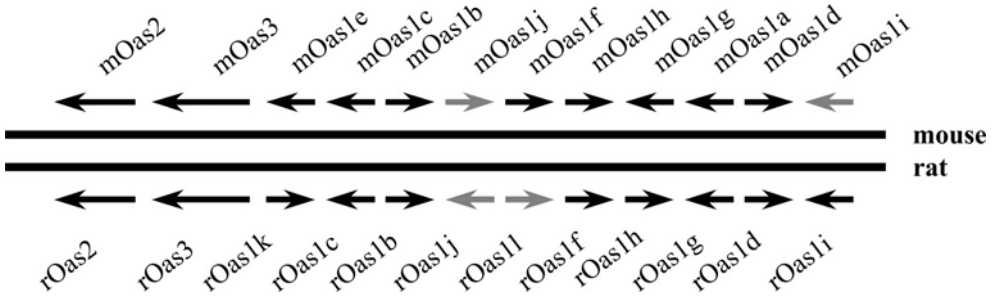


Figure 1. Structures of mouse and rat 2-5A synthetase gene families. Functional genes and pseudogenes are indicated as black and grey arrows, respectively.

A new pig pOAS1Y gene and two new cattle genes, cOAS1Y and cOAS1Z, were also indentified. Exact orthologous relationships between cattle and pig OAS1 genes could not be established due to the lack of a pig OAS1 map. New and previously reported sequences from even-toed ungulates and rodents were used to build a phylogenetic tree (Fig. 2). Two parts of the tree are in discordance with the map of the rodent Oas gene cluster. First, the rodent Oas1c and Oas1d genes are separated from each other on the map by six other genes and show a good orthological correspondence between mouse and rat. However, comparison of their sequences showed that the differences between paralogous rodent Oas1c and Oas1d genes in each species are smaller than those between the corresponding orthologs. Second, the rOas1g gene is more similar to rOas1i than to mOas1g. Therefore, in two cases, the similarities between rodent Oas1 orthologous cDNA sequences are lower than those between Oas1 paralogous sequences within each species. The observed deviation from the principle of divergent evolution, which assumes that more recent gene duplication produces more similar sequences, supports concerted

evolution of paralogous Oas1 genes in rodents presumably due to gene conversion. Even in cases where the tree topology was correct, the effect of gene conversion resulted in non-uniform occurrence of convergent substitutions. There were 151 and 159 convergent substitutions between paralogous genes in mice and rats, respectively, whereas only 111 convergent substitutions were observed between non-orthologous rodent Oas1 genes. Comparison of the substitution distributions along the nucleotide sequence in different evolutionary lineages can also unambiguously demonstrate the presence of gene conversion. A new approach was developed to quantify the differences between distributions of substitutions. We assumed that the substitution rate at any particular nucleotide position does not change among the genes and species being considered. However, different nucleotide positions may have different substitution rates. For each pair of genes, the cumulative distribution of differences can be built by assigning to each position i the number of differences N_i located to the left of this position normalized by the total number of differences N observed between the two genes, i.e., $F_i = N_i/N$. The distributions for two pairs of sequences are then compared by a non-parametric Kolmogorov-Smirnov (KS) test, which estimates the maximum difference between the two distributions. The distributions are expected to be similar when there are no evolutionary events or restrictions differentially affecting the occurrence of sequence changes in different lineages. For the cOAS1Y/cOAS1Z pair of genes, the KS test was not significant (P = 0.512). The cOAS1X/cOAS1Z pair showed a higher substitution rate (P = 0.069) in the proximal half of the gene compared to the distal half. For the cOAS1X/cOAS1Y pair, the result of the KS test was highly significant ($P = 4.9 \times 10^{-9}$). This pair has a low substitution rate in the first half of the sequence and a significantly higher rate in the second half. For the cOAS1X/cOAS1Z and cOAS1X/cOAS1Y pairs, the point at which the substitution rates switch is located at approximately the same position, namely position 500 of the cOAS1X coding sequence (difference 0.47). Therefore, cOAS1X is more similar to cOAS1Y before position 500, but cOAS1X is more similar to cOAS1Z after this position. The simplest explanation of this observation is that cOAS1X originated by conversion between the cOAS1Z and cOAS1Y genes and then diverged. A similar analysis was performed for the pig genes, pOAS1X and pOAS1Y. The distribution of differences between these two genes reached 0.47 at position 940 ($P = 1.4 \times 10^{-9}$), which could be due to a recent transfer of a large portion from one gene to another. The presence of interlocus transfers between duplicated copies of genes violates the basic principle of phylogenetic inference, i.e. the principle of divergence of related sequences, suggesting that gene conversion is the major mechanism by which homogenization occurs in paralogous 2'-5' oligoadenylate synthetase-1 genes.

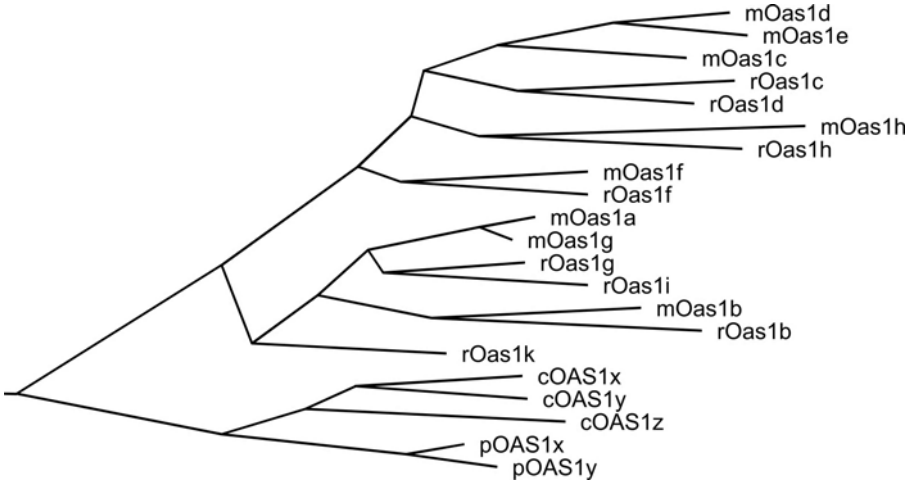


Figure 2. A phylogenetic tree of mouse (m), rat (r), cattle (c) and pig (p) Oas1 genes.

ACKNOWLEDGEMENTS

This work was supported by Public Health Service grants AI045135 from the Institute of Allergy and Infectious Diseases, National Institutes of Health and CI000216 from the National Center for Infectious Diseases, Centers for Disease Control and Prevention.

REFERENCES

Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W., Lipman D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res*, **25,** 3389–3402.

Justesen J., Hartmann R., Kjeldgaard N.O. (2000) Gene structure and function of the 2'-5'-oligoadenylate synthetase family. *Cell Mol. Life Sci*., **57,** 1593–1612.

Perelygin A.A., Lear T.L., Zharkikh A.A., Brinton M.A. (2005) Structure of equine 2'-5' oligoadenylate synthetase (Oas) gene family and FISH mapping of Oas genes to ECA8p15-p14 and BTA17q24-25. *Cytogenetics and Genome Research*, **111,** 51–56.

Perelygin A.A., Scherbik S.V., Zhulin I.B., Stockman B.M., Li Y., Brinton M.A. (2002) Positional cloning of the murine flavivirus resistance gene. *Proc. Natl. Acad. Sci. USA*, **99,** 9322–9327.

Zharkikh A., Li W.H. (1995) Estimation of confidence in phylogeny: the complete and partial bootstrap technique. *Mol. Phylogenet Evol*., **4,** 44–63.

POPULATION GENETIC POLYMORPHISM OF ENDEMIC MOLLUSCS *BAICALIA CARINATA* (MOLLUSCA: CAENOGASTROPODA)

Peretolchina T.E.*[*], *Bukin Yu.S.*, *Sitnikova T.Ya.*, *Sherbakov D.Yu.
Limnological Institute, SB RAS, Irkutsk, Russia
[*] Corresponding author: e-mail: tanya@lin.irk.ru

Key words: *Baicalia carinata*, mtCO1, genetic polymorphism, *U*-criterion of Mann and Whitney

SUMMARY

Motivation: Sets of aligned nucleotide sequences resulting population level genetic polymorphism studies are generally believed to contain more ecologically and evolutionary meaningful information than is uncovered by phylogenetic analysis. This paper aims at the elucidation of population structure of Baikalian endemic snail *Baicalia carinata.*

Results: Basing on 97 nucleotide sequences of the 588 base pairs long fragment of mitochondrial gene coding for 1 subunit of cytochrome *c* oxydase of Baikalian endemic gastropod *Baicalia carinata* (Dybowski, 1875) we show that this species is split into three clades, which correspond to three populations: Southern, North-Western and North-Eastern. We used the Mann and Whitney rank criterion in order to estimate the statistic significance of genetic differences between the populations.

INTRODUCTION

Baicalia catinata belongs to endemic family Baicaliidae from Lake Baikal. High intra-specific shell polymorphism is peculiar to this family and makes morphological traits based species identification a very difficult task. During more then century-long study of baikalian mollusks their taxonomy was many times revised basing on morphological traits: structure of shell and radula (Dybowski, 1875; Lindholm, 1909; 1924) and the anatomy of female sexual system (Sitnikova, 1991). Grochmalicki and Dybowski (1914) described 10 intraspecific forms of *B. carinata*, which were considered to be ecological forms by Kozhov (1936) of the same species. Here we use molecular phylogeny of mitochondrial cytochrome *c* oxidase subunit 1 (mtCO1) gene nucleotide sequences in order to study population genetic structure of *B. carinata*.

METHODS

Mollusks were collected by dredge at 11 localities of Lake Baikal at a depth of 10–40 meters. After preliminary sorting of a benthic sample, mollusks were fixed in 80 % ethanol for 24 hours with subsequent ethanol change with the 70 % solution and kept until the extraction of DNA.

Genomic DNA was extracted by modified methods of Doyle and Dikson from muscle tissue (Doyle, Dickson, 1987).

The mtCO1 fragment was amplifed DNA with the PCR method using universal primers for invertebrates (Folmer *et al*., 1994). Sequences were determined by DNA

analyse system CEQ 8800 (Beckman Coulter Inc) using the same primers. All obtained sequences were aligned relatively of *B. carinata* sequence from GenBank (Accession Number Z92993), using computer program BioEdit ver.5.0.9.

Neighbor Joining phylogenies were obtained for analysis intra-specific polymorphism.

Criterion of Mann and Whitney was used in order to estimate the statistic significance of genetic differences between populations (Mann, Whitney, 1947). *U* verifies null-hypothesis: two independent samples belongs the identical general totality and their distribution function of probabilities equals i.e. $F_1(x) = F_2(x)$. In our case *x* is number of substitutions (percent) between nucleotide sequences. In order to calculate statistic *U* it is necessary to sort $(m+n)$ values of the combined sample and to rank them. The sum of ranks for first population is R_1, for the second population it is R_2. U_1 and U_2 are calculated as follows:

$$U_1 = mn + \frac{m(m+1)}{2} - R_1$$

$$U_2 = mn + \frac{n(n+1)}{2} - R_2,$$

where *m* is number of ranks in population A, *n* is number of ranks in sample B. The smaller one of U_1 and U_2 is then chosen as the value of the criterion. Null-hypothesis is rejected if U -value is less then the critical value or equal to it. If size of samples not too small ($m \geq 8$, $n \geq 8$), we still can use the following equation to estimate statistic significance of null-hypothesis:

$$\hat{z} = mn + \frac{\left| U - mn/2 \right|}{\sqrt{\frac{mn(m+n+1)}{12}}}, \quad (1)$$

where $\hat{z}$ is value determining the statistic significance of the difference between the samples. In case of the populations significantly different, one may use this value to quantify the difference.

U-criterion is $mn/2$ in case if the populations are identical, and approaches 0 with the increase of differences between them. Therefore the amount of difference between the populations K_0 may be calculated as follows:

$$K_0 = \left(1 - \frac{2U}{mn}\right), \quad (2)$$

K_0 varies from 0 to 1.

This coefficient can be interpreted in per cent value. Thus K_0 behaves like the F_{st} criterion from DNA sequence data (Hudson *et al*., 1992; Slatkin, 1995): it increases with the increase of the differences between populations.

RESULTS AND DISCUSSION

In total, 97 sequences of mtCO1 fragment were obtained for 97 specimen of *Baicalia carinata* from 11 localities of Lake Baikal. The length of the fragment sequenced is 588 b.p. All sequences were deposited in GenBank with the Accession Numbers from DQ436347 to DQ436443.

Genetic polymorphism between nucleotide sequences varies from 0 to 4.3 % within one locality and may be up to 6.8 % between sequences from the most distant localities.

The phylogenetic NJ tree shows that *B. carinata* form three clades. Specimen of these clades occupies geographically continuous areals. The first clade includes mollusks from Chivirkui Bay, near Gremyachinsk town, P. Pongonie, Isl. Listvenicnyi and P. Tonkyi (North-East population). Representatives of the second clade inhabit Bolshie Coty and Murinskaya Bank (South population). The third clade consists of snails from Olchon Gate Strait, Isl. Yarki and Tutai Bay (North-West population).

In order to estimate the statistic significance of difference between populations, the rank criterion of Mann and Whitney (*U*-criterion) was used. This criterion was chosen because it's performance does not depend on the shapes of distributions compared while F_{st}-criterion is correct only in case of Gaussian distributions. This method is simple and does not require large amount of computations.

We calculated $\hat{z}$ value, which determines significance of differences between samples by equation (1). The value obtained suggests that populations (North-East, South and North-West) are distinctive with 99 % probability.

With equation (2) we calculated coefficient of separation - K_0. The result of analysis shows that North-East and North-West populations separated by 21.5 %, North-West and South populations are separated by 33.2 %, and North-East and South populations are 14.7 % different (Fig. 1). Therefore we can conclude that the least exchange by genetic information takes place between South and North-West populations, and South and North-East populations are the closest.

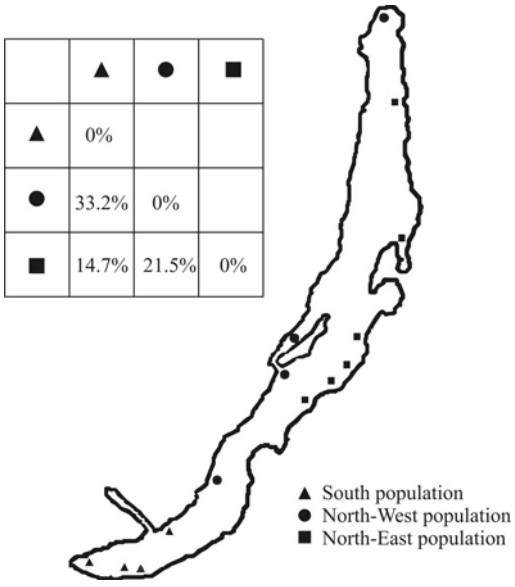


Figure 1. Map of Lake Baikal, indicating localities of collecting samples and matrix of population disjoining in percent.

REFERENCES

Dybowski W. (1875) Die Gastropodenfauna des Baikal-Sees, Anatomisch und sistematisch bearbeited. *Memories de l'Academie Imperiale des Sciences de St. Petersbourg*, **22**(8), 73.

Dybowski B., Grochmalicki J. (1914) Beiträge zur Kenntnis der Baikalmollusken. *Buchdruckerei der keiserlichen Academie der Wissenschaften, Petrograd*, **19**(2), 286–322.

Doyle J.J., Dickson E. (1987) Preservation of plant samples for DNA restriction endonuclease analysis. *Taxon*, **36,** 715–722.

Folmer O., Black M., Hoeh W., Lutz R., Vrijenhoek R. (1994) DNA primer for amplification of mitochondrial cytochrome *c* oxidase subunite I from diverse metazoan invertebrates. *Mol. Marine Biol. and Biotechnology*, **3,** 294–299.

Hudson R.R., Slatkin M., Maddison W.P. (1992) Estimation of levels of gene flow from DNA sequence data. *Genetics*, **132,** 583–589.

Kozhov M.M. (1936) Mollusks of Lake Baikal. Moscow, RAS, 320 p.

Lindholm W.A. (1924) Enige neue Gastropoden aus dem Baikalsee. *Proceedings of Russian Academy Science*, 22–25.

Lindholm W.A. (1909) Die Mollusken des Baikal-Sees (Gastropoda et Pelycopoda). Systematisch und zoogeographisch bearbeitet. Wissenschaftliche Ergebnisse einer Zoologischen Expedition nacht dem Baikal-See unter Leitung des professors alexis Korotneff in den Jahren 1900–1902. *Zoological investigation of Lake Baikal, Kiev*, **4,** 104.

Mann H.B., Whitney D.R. (1947) On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Statist.*, **18,** 50–60.

Sitnikova T.Ya. (1991) The new structure of endemic baikalian family *Baicaliidae* (Mollusca, Gastropoda, Pectinibranchia). *Collected works of Limnological Institute RAS*, 281–295.

Slatkin M.A. (1995) Measure of population subdivision based on microsatellite allele frequencies. *Genetics*, **139**(1), 457–462.

PHYLOGENETIC ANALYSIS OF THE p53 AND p63/p73 GENE FAMILIES

***Pintus S.S.*[*1], *Ivanisenko V.A.*[2]**
[1] Novosibirsk State Unversity, Novosibirsk, 630090, Russia; [2] Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia
[*] Corresponding author: e-mail: pintus@bionet.nsc.ru

Key words: p53, p63/p73, phylogenetic analysis, multiple alignment, positive selection, purifying selection

SUMMARY

Motivation: Proteins of the relative families p53 and p63/p73 are transcriptional factors that are involved in the signaling pathway in cells. The wide spectrum of their functions includes cell cycle arrest and apoptosis in response to DNA damage. The p53 protein also participates in development of particular tissues during embryogenesis. Thus, it is of high importance to establish the relation between structure, function and evolution of these proteins.

Results: In the current computational study, the evolutionary mode of the p63/p73 protein family was investigated. The results obtained were compared with our previous of the phylogenetic analysis of the p53 protein. Evidence was obtained, indicating that the evolutionary history of the p63/p73 proteins has been under positive selection. An attempt was made to associate the current evidence with the previous for positive selection in the p53 family.

INTRODUCTION

The p53 protein is an important tumor suppressor because of its ability to produce apoptosis in the tumor cells. This protein can also arrest the cell cycle and allows cell repair before the start of replication. The p53 protein also plays an important role in the embryonic development of vertebrates by regulating the proliferation and apoptosis of the cells (Levine *et al*., 2004). The p53 protein is a relative of the p63/p73 protein family, which is somewhat structurally similar to p53, but differs from it by the presence of the C-terminal domain. The proteins of this family can induce cell cycle arrest and apoptosis, but they are of greater importance in embryogenesis than p53. Establishment of the relation between structure, function and evolution of these two families is of importance because it would provide a better understanding of the mechanism of their antitumor activity and involvement in embryogenesis (Saccone *et al*., 2002).

The aim of this study was to define the relation between the functional significance of particular amino acid residues of the p63/p73 protein family, their position in the protein structure, and the evolutionary mode of the codons, that correspond to the residues. Another aim was to compare the current results with those previously obtained for the p53 family (Benson *et al*., 2006).

METHODS AND ALGORITHMS

Amino acids and the nucleotide sequences for the p63/p73 proteins of 9 vertebrate species were taken from the GenBank database (Benson *et al*., 2006) like in our previous study for the p53 protein (Pintus *et al*., 2006). The species and accession numbers of the corresponding database entries are listed in Table 1. We obtained a multiple alignment of the amino acid sequences using the ClustalW program, version 1.7 (Li *et al*., 2003) applied to build the phylogenetic tree using the PHYML (Guindon *et al*., 2003) package, version 2.4.4.

Also, a multiple nucleotide alignment was obtained on the basis of the residue alignment using the *ad hoc* Perl program. With this program, instead of amino acid residues, their corresponding codons from the nucleotide sequences of the corresponding genes were inserted in the alignment text. Accordingly, each gap of the alignment was replaced by 3 gaps of the nucleotide alignment. Search for the adaptive branches of the phylogenetic tree and the adaptive codons in the nucleotide sequences was performed using the codeml program from the PAML package, version 3.14 (Yang *et al*., 2000).

Table 1. Accession numbers of the GenBank entries for p63 and p73 proteins in different species

p63		p73	
NM_011641	*Mus musculus*	Y19234	*Mus musculus*
NM_019221	*Rattus norvegicus*	XM_342992	*Rattus norvegicus*
NM_204351	*Gallus gallus*	XM_417545	*Gallus gallus*
NM_003722	*Homo sapiens*	NM_005427	*Homo sapiens*
XM_845322	*Canis familiaris*	XM_546740	*Canis familiaris*
BC076530	*Danio rerio*	NM_183340	*Danio rerio*
AF314148	*Xenopus laevis*	AF043641	*Barbus barbus*
XM_867115	*Bos Taurus*	XM_593064	*Bos taurus*
XM_516946	*Pan troglodytes*	Y11419	*Cercopithecus aethiops*

IMPLEMENTATION AND RESULTS

It was found that p73 evolution was associated with positive selection at the time when the Carnivora and Artiodactyla ancestors diverged from each other. The codon that predominantly accumulated nonsynonymous substitutions was detected in the coding sequence of the p53 gene. In the human p53 protein this codon occupies position 556T.

Evidence was obtained indicating that during the evolutionary history of the p63 protein positive selection acted at the time of divergence between Synapsida and Reptilia (dN/dS = 1.9352). The dendrograms for both p73 and p63 proteins are depicted in Fig. 1 and 2, respectively.

DISCUSSION

Our previous study demonstrated that positive selection during the divergence of Synapsida an Amniota, in general, acted also throughout the evolutionary history of the p53 protein, a relative of the p63/p73 protein family. It is known that homoiothermy and viviparity first appeared among synapsids. These evolutionary acquirements have made necessary changes in the genetic regulation of ontogeny, and this, in turn, might have caused adaptive changes in the p53 and p63/p73 families. However, the appearance of the terrestrial animals and the emergence of homoiothermy demanded a higher metabolic rate that most likely acted as a carcinogen.

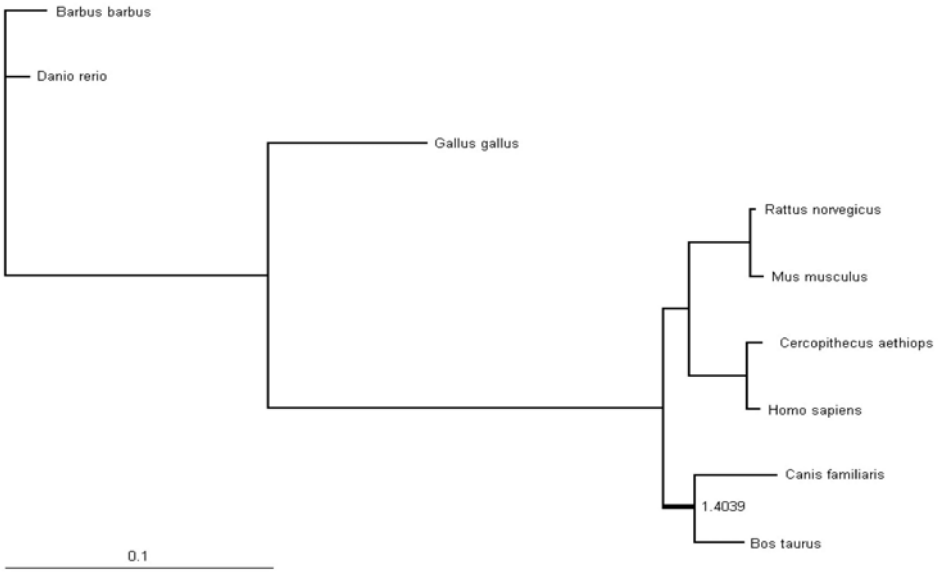


Figure 1. Dendrogram of p73 protein. The branch that corresponds to positive selection is shown in bold and the dN/dS ratio is indicated.

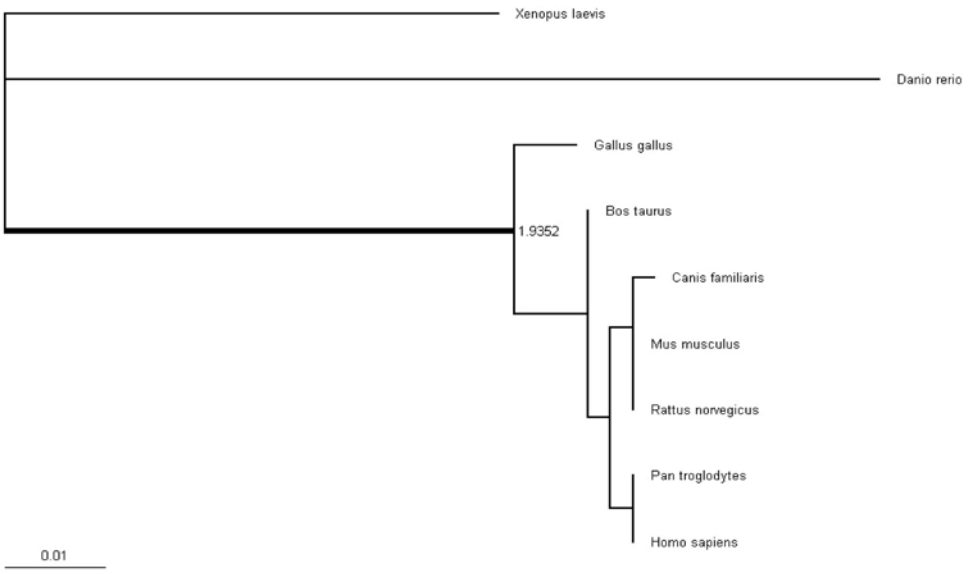


Figure 2. Dendrogram of p63 protein. The branch that corresponds to positive selection is shown in bold and the dN/dS ratio is indicated.

These conclusions are consistent with the idea that the p53 protein has descended from an ancestor common to p63/p73, and it has lost during its evolution the C-terminal SAM domain at about the time of the divergence of teleosts and amphibia (Saccone *et al*., 2002). In the meantime, p53 function underwent profound specialization possibly because of the increase in the rate of tissue regeneration in the advanced vertebrates leading to higher risk of tumor formation (Yang *et al*., 2002).

ACKNOWLEDGEMENTS

The authors are grateful to I.V. Lokhova for bibliographical support and A.N. Faddeva for translation of the abstract into English. The work was supported in part by Russian Foundation for Basic Research (Nos 04-01-00458, 05-04-49283, 06-04-49556), the State contract No. 02.434.11.3004 of 01.04.2005 and No. 02.467.11.1005 of 30.09.2005 with the Federal Agency for science and innovation "Identification of promising targets for the action of new drugs on the basis of gene network reconstruction" federal goal-oriented technical program "Study and development of priority directions in science and technique, 2002–2006, Interdisciplinary integrative project for basic research of the SB RAS No. 115 "Development of intellectual and informational technologes for the generation and analysis of knowledge to support basic research in the area of natural science", CRDF Rup2-2629-NO-04.

REFERENCES

Benson D.A., Karsch-Mizrachi I., Lipman D.J., Ostell J., Wheeler D.L. (2006) GenBank. *Nucl. Acids Res*., **34**, D16–D20.

Guindon S., Gascuel O. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol*., **52**, 696–704.

Levine A.J., Finlay C.A., Hinds P.W. (2004) P53 is a tumor suppressor gene. *Cell*, **116**, S67–S69.

Li K.B. (2003) ClustalW-MPI: ClustalW analysis using distributed and parallel computing. *Bioinformatics*, **19**, 1585–1586.

Pintus S.S., Fomin E.S., Ivanisenko V.A., Kolchanov N.A. (2006) Phylogenetic analysis of the p53 family. *Biofizika* (in press).

Saccone C., Barome P.O., D'Erchia A.M., D'Errico I., Pesole G., Sbisa E., Tullo A. (2002) Molecular strategies in Metazoan genomic evolution. *Gene*, **300**, 195–201.

Yang Z., Bielawski J.P. (2000) Statistical methods for detecting molecular adaptation. rends. *Ecol Evol*., **15**, 496–503.

Yang A., Kaghad M., Caput D., McKeon F. (2002) On the shoulders of giants: p63, p73 and the rise of p53. *Trends Genet*., **18**, 90–95.

HUMAN GENOME POLYMORPHISM AND ALTERNATIVE SPLICING

***Ramensky V.*[*1], *Nurtdinov R.*[2], *Neverov A.*[3], *Mironov A.*[2], *Gelfand M.*[2, 4]**
[1] Engelhardt Institute of Molecular Biology, RAS, Moscow, 119991, Russia; [2] Department of Bioengineering and Bioinformatics, Moscow State University, Moscow, 119992, Russia; [3] State Scientific Center GosNIIGenetika, Moscow, 117545, Russia; [4] Institute for Information Transmission Problems, RAS, Moscow, 127994, Russia
[*]Corresponding author: e-mail: ramensky@imb.ac.ru

Key words: alternative splicing, single nucleotide polymorphism, McDonald-Kreitman test, positive selection, molecular evolution

SUMMARY

Motivation: Alternative splicing and single nucleotide polymorphism are the phenomena responsible for the organismal molecular diversity. We present the study of polymorphism in the coding regions of 9,125 alternatively spliced human genes aimed at understanding of the forces driving their evolution.

Results: The significant difference in non-synoymous to synonymous SNP ratio is observed: 0.757 in constitutive exons vs. 1.799 in alternative ones from minor isoforms. The alternatively spliced regions therefore experience lower selective pressure at the amino acid level. The results of the McDonald-Kreitman test suggest that, unlike the constitutive regions, they are also subject to the positive selection.

INTRODUCTION

Alternative splicing (AS) of genes has been recognized as one of the major sources of organismal complexity (Graveley, 2001). The AS can be defined as the various ways of splicing out introns in eukaryotic pre-mRNAs resulting in one gene producing several different mRNAs and protein products (isoforms). The estimates of fraction of alternatively spliced genes in the human genome gradually increase from the level of 35 % (Mironov, 1999) up to 70–80 % (Johnson, 2003; Kampa, 2004). The alternative splicing has also been recognized as the mechanism of accelerated evolution by relaxation of purifying selection pressure (Cusack, Wolfe, 2005; Xing, Lee, 2005; Ermakova *et al*., 2006, in press). Single nucleotide polymorphism (SNP) is another well-recognized phenomenon providing molecular diversity (Brookes, 1999) and comprising approximately 90 % of human DNA variation (Collins *et al*., 1998). The SNPs in the coding regions of genes fail in two categories, synonymous or silent (sSNPs) and non-synonymous (nsSNPs) that change the corresponding amino acid residue and therefore are thought to be responsible for existence of various phenotypes (Collins *et al*., 1998). The relationship between the non-synonymous (Pa) and synonymous (Ps) SNP (Zhao *et al*., 2003) or fixed interspecies variation (Da, Ds) (Xing, Lee, 2005) provides information on selection pressure in the genomic regions under study. The analysis of SNP density in the Celera human sequence assembly showed that the bulk genomic ratio Pa/Ps is less than half of that under neutral expectations, reflecting the acting purifying (negative) selection (Zhao *et al*., 2003). The simultaneous account for the four abovementioned types of variation (Da, Ds, Pa, Ps) enables the detection of positive darwinian selection in

the presence of negative selection via the McDonald-Kreitman test (McDonald, Kreitman, 1991). In this work, we investigate the differences between patterns of synonymous and non-synonymous polymorphism and human-chimpanzee variation in alternative and constitutive regions of human genes. The novelty of the approach is in the simultaneous analysis of polymorphism and divergence data in the coding regions of two types that reveals not only the relaxation of selection pressure (Cusack, Wolfe, 2005; Xing, Lee, 2005) but also the significant difference between the positive selection strength in the alternative and constitutive regions.

METHODS AND ALGORITHMS

Validated SNPs from build 121 of the dbSNP database (Sherry *et al*., 2001) were mapped to human genes from the EDAS database (Nurtdinov, 2004) with at least two isoforms generated by the IsoformCounter algorithm (Neverov *et al*., 2005). The human-chimpanzee synonymous and non-synonymous variation were derived from the whole genome alignments (Kent *et al*., 2002). The conservative isoform generation procedure and subsequent postprocessing (regions between the isoform ends and stops, alternatives with frameshifts, and regions not completely aligned to the chimp genome are not considered) generates 64,742 constitutive and 18,036 alternatively spliced regions from 9,125 genes (Table 1). The considered regions include cassette exons, alternative exons, and alternative 5′ and 3′ splice sites.

Table 1. Alternative and constitutive regions. Minors are those included in less than 2/3 of all coding sequences (ESTs, mRNAs, peptides) observed for this regions of a gene

	Num. of regions	Total length
Constitutive	64,742	9,544,015
Major	11,129	1,193,395
Minor	6,907	720,845

IMPLEMENTATION AND RESULTS

The resulting statistics on intra- and interspecies variation is given in the Table 2. The significant difference between Pa/Ps ratio in constitutive and alternative gene regions is observed. The data on human-chimp variation follows the same trend.

Table 2. SNPs and human-chimp mismatches in the various regions of alternatively spliced genes

	Pa = SNP Nonsyn	Ps = SNP Synon	Pa/Ps	Da = Mismatch Nonsyn	Ds = Mismatch Synon	Da/Ds
Constitutive	4380	5787	0.757	25574	34104	0.750
Major	588	707	0.832	3164	4268	0.741
Minor	644	358	1.799	4687	2377	1.972

However, the excess of nonsynonymous substitutions relative to polymorphism (Da/Ds = 1.972 > Pa/Ps = 1.799) in minor alternative regions implies the fixation of advantageous mutations (Fay *et al*., 2002). This effect increases if the most stringent conditions, when only mRNA and protein sequences are considered: Da/Ds = 1.892 vs. Pa/Ps = 1.659 (data not shown). The P-value of the χ^2-test is in the range 0.1–0.2. This value can be explained by the relatively small number of SNPs.

DISCUSSION

The excess of non-synonymous SNPs and human-chimp mismatches in the alternatively spliced regions of human genes suggests that they are under lower purifying selective pressure. Unlike the constitutive regions, they also experience the positive selection, as shown by the results of McDonald-Kreitman test. The simultaneous action of these two forces shapes the evolution of alternatively spliced genes.

ACKNOWLEDGEMENTS

Authors thank E. Ermakova for valuable discussions. This work was supported by the grants "Human molecular polymorphism" and "Cellular and Molecular Biology" from the Russian Academy of Sciences.

REFERENCES

Brookes A.J. (1999) The essence of SNPs. *Gene*, **234**, 177–186 (review).

Collins F.S. *et al*. (1998) A DNA polymorphism discovery resource for research on human genetic variation. *Genome Res.*, **8**,1229–1231.

Cusack B.P., Wolfe K.H. (2005) Changes in alternative splicing of human and mouse genes are accompanied by faster evolution of constitutive exons. *Mol. Biol. Evol.*, **22**, 2198–2208.

Graveley B.R. (2001) Alternative splicing: increasing diversity in the proteomic world. *Trends Genet*, **17**, 100–107 (review).

Ermakova E.O. *et al*. (2006) Faster evolutionary rate in alternatively spliced regions. *BMC Genomics* (in press).

Fay J.C. *et al*. (2002) Testing the neutral theory of molecular evolution with genomic data from Drosophila. *Nature*, **415**, 1024–1026.

Johnson J.M. *et al*. (2003) Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science*, **302**, 2141–2144.

Kampa D. *et al*. (2004) Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Res*., **14**, 331–342.

Kent W.J. *et al*. (2002) The Human Genome Browser at UCSC. *Genome Res*., **12,** 996–1006.

Mironov *et al*. (1999) Frequent alternative splicing of human genes. *Genome Res.*, **9**, 1288–1293.

McDonald J.H., Kreitman M. (1991) Adaptive protein evolution at the Adh locus in Drosophila. *Nature*, **351**, 2–4.

Neverov A.D. *et al*. (2005) Alternative splicing and protein function. *BMC Bioinformatics*, **6**, 266.

Nurtdinov R.N. *et al*. (2004) EDAS - EST-Derived Alternative Splicing Database. *Nucl. Acids Res.*, *Mol. Biol. Database Collectionentry*, 631.

Sherry S.T. *et al*. (2001) dbSNP: the NCBI database of genetic variation. *Nucl. Acids Res.*, **29**, 308–311.

Xing Y., Lee C. (2005) Evidence of functional selection pressure for alternative splicing events that accelerate evolution of protein subsequences. *Proc. Natl. Acad. Sci. USA*, **102**, 13526–13531.

Zhao Z. *et al*. (2003) Investigating single nucleotide polymorphism (SNP) density in the human genome and its implications for molecular evolution. *Gene*, **312**, 207–213.

PRIMORDIAL TRANSLATION OF BOTH COMPLEMENTARY STRANDS MIGHT DIRECT THE EARLY EVOLUTION OF THE GENETIC CODE

***Rodin S.N.*[*1], *Rodin A.S.*[2]**
[1] Beckman Research Institute of the City of Hope, Duarte, CA 91010, USA; [2] School of Public Health, University of Texas, Houston, TX 77225, USA
[*] Corresponding author: e-mail: srodin@coh.org

Key words: RNA world, the genetic code, double-strand coding, tRNA, aaRS

SUMMARY

Motivation: In the bilingual world of nucleic acids and proteins, usually only one of the gene's two DNA strands, namely a sense strand, is protein-coding. The preceding monolingual RNA world was likely much more strand-symmetric so that both (plus and minus) complementary replicas of ancestral genes could have been used with equal success as first templates (Rodin, Ohno, 1995, 1997; Carter, Duax, 2002). One would expect then for this primordial strand symmetry to be reflected in the genetic code organization.

Results: The updated analysis of the current tRNA gene compilation (Sprinzl, Vassilenko, 2005) confirmed and strengthened our previous discovery that in pairs of consensus/ancestral tRNAs with complementary anticodons, their 2nd bases in the acceptor stem are also complementary (Rodin *et al*., 1996; Rodin S., Rodin A., 2006). Here we show that, indeed, it might have been an in-frame double-strand coding of RNA genes that has directed the code's earliest expansion and preserved this fundamental complementary link between the acceptors and the anticodons.

Availability: All datasets used in this study and the reconstructed tRNA trees are available from the authors. No proprietary software has been used in this study.

INTRODUCTION

If the origins of the genetic code are still on the record anywhere, the best candidates for finding vestiges of this record are the two main code adapters, tRNAs and aminoacyl-tRNA synthetases (aaRS). Two sites in the tRNA molecule are responsible for bringing the code into action, the 3′ end of the acceptor stem to which the specific amino acid is attached by the cognate aaRS, and the anticodon that determines this specificity. It thus appears that there are actually two codes – the classic code represented by anticodons for reading codons in mRNA, and the "second" (De Duve, 1988) operational code (Schimmel *et al*., 1993; Schimmel, 1996) localized mainly in the acceptor for correct aminoacylation at its 3′end. However, in L-shaped fold of tRNAs these two sites are separated by the largest physical distance topologically possible, ~75Å. How could they "communicate" in ancient tRNAs without aaRSs?

Previously we have found that in pairs of consensus tRNAs with complementary anticodons their bases at the 2nd position of the acceptor stem were also complementary (Rodin *et al*., 1996). In fact, this dual complementarity is a still-preserved vestige of the common evolutionary root for the two codes. Specifically, it suggests that the anticodon

triplet and the first three bases of the acceptor stem were, at the very beginning, one and the same (Rodin *et al*., 1996; Rodin, Ohno, 1997). However, the dual complementarity *per se* does not address the crucial question: When did the operational code originate and how could it co-evolve with the expanding set of amino acids and anticodons? The present study aims to provide an answer.

MATERIAL AND METHODS

In total, we have analyzed 8,246 tRNA gene sequences from the latest Genomic tRNA Compilation (Sprinzl, Vassilenko, 2005) covering eubacteria, archaebacteria and eukaryotes. We have manually reconstructed the ancestral tRNA sequences, separately for each anticodon, for each of these three kingdoms and for their common ancestor. (We used MEGA 3.1 phylogenetic analysis software (http://www.megasoftware.net/) to build the tRNA trees; the tree topologies proved sufficiently robust for our purposes with respect to the phylogenetic reconstruction method and substitution model used.) As a rule, the ancestral 2nd bases coincided with the consensus ones.

As a measure of the dual complementarity, DC, we used the number of tRNA pairs with complementary anticodons in which the 2nd bases of the acceptor helix were also complementary, divided by the total number of such pairs. The DC values have been calculated not only for strictly complementary anticodons with canonical G–C and A–U base pairings, but also for the anticodons with illegitimate R–Y complementarity: the G:U "wobbling" pairing and even A*C mispairing. Since some of reconstructed ancestral (consensus) nucleotides in the acceptor helix were uncertain, we calculated the DC values for the two extreme cases – first, these uncertainties were all counted against the dual complementarity thus yielding the DC_{min} index; second, these uncertainties were all counted in a favor of the dual complementarity thus yielding the DC_{max} value (Rodin S., Rodin A., 2006). The real DC values lie somewhere in between.

RESULTS AND DISCUSSION

1. The main result is that the dual complementarity is shown by ancestral / consensus tRNA pairs with completely complementary anticodons and is not shown by tRNA pairs in which only the 2nd bases of anticodons are complementary. The corresponding average DC indexes are 0.74(min) – 0.86(max) (total 69 pairs, flanking wobbling R-Y pairings included) versus 0.38(min) – 0.49(max) (total 214 pairs), respectively. This difference indicates that although from the very beginning the operational code probably used the double-stranded anticodon(codon)-like triplets in the acceptor stem, this primordial code was quite ambiguous, i.e. only the 2nd base pair was important at the time when the first protein aaRSs began to replace their iso-specific ribozymic precursors (for more detail see: Rodin S., Rodin A., 2006). Furthermore, the following two corollaries immediately come to mind:

By the time ancestral tRNAs gained the dual complementarity, the 3-letter translation frame had already been in use, and

The very phenomenon of dual complementarity is possible because the new tRNAs entered primitive translation in pairs with complementary anticodons.

Obviously, the same holds for corresponding codons and amino acids. Simultaneous involvement of tRNAs with complementary anticodons in translation is easily realized by in-frame double-strand coding (Rodin, Ohno, 1995, 1997; Rodin *et al*., 1996). Fig. 1 illustrates this for a gene containing, for simplicity, only two complementary triplets, GGC and GCC, encoding Gly and Ala, respectively.

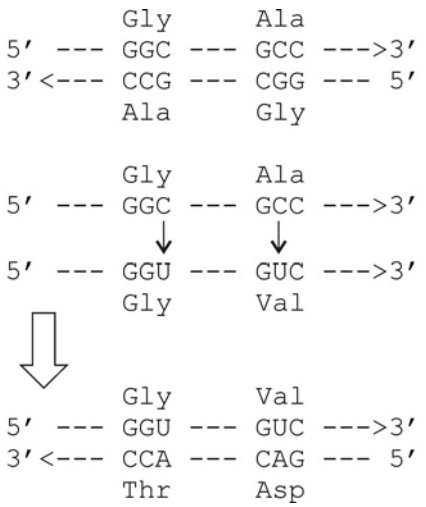


Figure 1. Example of code expansion by translation of both complementary strands.

Both strands serve as templates for translation. Spontaneous C→U transition in one strand, in the center of GCC codon, produces a new triplet GUC that specifies Val, necessarily entailing the GGC→GAC mutation in the opposite strand that specifies Asp (Fig. 1). Clearly, the concerted recruitment of these new amino acids in the genetic coding machinery implies at least two duplication events for $tRNA^{Ala}$ and $tRNA^{Gly}$ genes, with their subsequent mutational "tunes-up". Also shown is an example of C:G→U:A transitions occurring in flanking 1st and 3rd codon positions usually complemented by either synonymous or functionally conservative mutations on the opposite strand. Importantly, if the first tRNA couple (such as $tRNA^{Ala}$ and $tRNA^{Gly}$) had complementary 2nd bases in the acceptor stem, their duplicates with complementary mutated anticodons (such as $tRNA^{Val}$ and $tRNA^{Asn}$) could have maintained this dual complementarity while fixating more specific identity elements elsewhere in the tRNA molecule. This seems to be the case with (Ala/Gly → Val/Asp) or, for example, (Ala/Arg → Val/His) expansions of the codon repertoire (Rodin S., Rodin A., 2006).

2. The model of code expansion by double-strand coding predicts that tRNAs with G:U or A*C illegitimate pairings in the anticodons *will also show the dual complementarity*. Indeed, the illegitimate cases are just transitory mutational states in double-strand coding when one strand has already gained the mutation, whereas the opposite strand remains in the parental state, thus forming G:U or A*C irregularities. In reality, the *concerted* dual complementarity is observed with the flanking positions (DC = 0.77(min) – 0.89(max)), but this is certainly not the case with the central position (DC = 0.43(min) – 0.55(max)). Specifically, out of 16 amino acid tetrads *only two* appeared to be dual-complementary in *all four* combinations. These two are {Ala,Gly,Val,Asp}, and {Ala,Val,Arg,His}. The difference suggests that the 2nd nucleotide-based core structure of the genetic code was established for the above few (four to six) amino acids, with the subsequent expansion of the codon(anticodon)-to-aa assignment repertoire involving only the flanking functionally conservative, and even silent, nucleotides (Fig. 1). Remarkably, Gly, Ala, Asp and Val represent the most preponderant among abiotically synthesized amino acids (Miller, 1987). Another tetrad includes, again, Ala and Val, and two basic amino acids, Arg and His. Intriguingly, Arg and His show significant stereochemical affinity to cognate anticodons and/or codons in SELEX evolutionary tests (Yarus *et al*., 2005).

3. Consistent with the primacy of these two amino acid tetrads in most scenarios of the genetic code origin is the surprisingly distinct NRN/NYN bilateral pattern of the tRNA trees generated by the corresponding codons and their nearest mutational derivatives, which accords fully with the double strand coding-based expansion of the genetic code and still-

preserved dual complementarity. None of tRNA trees generated by other amino acid tetrads showed this pattern. The parallel phylogenetic examination of aaRSs is underway.

In conclusion, because of its original location in the acceptor helix and early expansion by archaic double strand translation, the code was not mutationally optimized for the subsequent dichotomization in sense and anti-sense strands (Rodin, Ohno, 1997). Based on this strand asymmetry, we propose a novel indicator of selection that is especially accurate for base transitions and transversions at palindromic CpG dinucleotides, and is expected to be particularly useful in distinguishing the selection- and drift-driven evolutionary pathways of duplicated genes in general, and duplicated genes encoding transcriptional factors and cofactors in particular.

ACKNOWLEDGEMENTS

We thank Paul Schimmel, Lluis Ribas de Pouplana and Germinal Cocho for discussions and valuable suggestions.

REFERENCES

Carter C.W., JR., Duax W.L. (2002) Did tRNA synthetase classes arise on opposite strands of the same gene? *Mol. Cell*, **10**, 705–708.

De Duve C. (1988) The second genetic code. *Nature*, **333**, 117–118.

Miller S.L. (1987) Which organic compounds could have occurred on the prebiotic earth. *Cold Spring Harbor Symp. Quant. Biol*., **52**, 17–27.

Rodin S., Ohno S. (1995) Two types of aminoacyl-tRNA synthetases could be originally encoded by complementary strands of the same nucleic acid. *Origins Life Evol. Biosphere*, **25**, 565–589.

Rodin S., Rodin A., Ohno S. (1996) The presence of codon-anticodon pairs in the acceptor stem of tRNAs. *Proc. Natl. Acad. Sci. USA*, **93**, 4537–3542.

Rodin S.N., Ohno S. (1997) Four primordial modes of tRNA-synthetase recognition determined by the (G,C) operational code. *Proc. Natl. Acad. Sci. USA*, **94**, 5183–5188.

Rodin S.N., Rodin A.S. (2006) Origin of the genetic code: First aminoacyl-tRNA synthetases could replace isofunctional ribozymes when only the second base of codons was established. *DNA Cell Biol*., **25**, 365–375.

Schimmel P. (1996). Origin of genetic code: A needle in the haystack of tRNA sequences. *Proc. Natl. Acad. Sci. USA*, **93**, 4521–4522.

Schimmel P., Giege R., Moras D., Yokoyama S. (1993) An operational RNA code for amino acids and possible relation to genetic code. *Proc. Natl. Acad. Sci. USA*, **90**, 8763–8768.

Sprinzl M., Vassilenko K.S. (2005) Compilation of tRNA sequences and sequences of tRNA genes. *Nucl. Acids Res*., **1**(33), D139–140.

Yarus M., Caporaso J.G., Knight R. (2005) Origins of the genetic code: The escaped triplet theory. *Annu. Rev. Biochem*., **74**, 125–151.

THEORETICAL ANALYSIS OF MITOCHONDRIAL DNA SOMATIC MUTATION SPECTRA IN OXYS AND WISTAR RATS

Rotskaya U.N., Rogozin I.B., Vasyunina E.A., Kolosova N.G., Sinitsyna O.I.*
Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia
* Corresponding authors: e-mail: olgasin1@yandex.ru

Key words: somatic mitochondrial mutations, mtDNA, OXYS rats

SUMMARY

Motivation: Analysis of mutations in mitochondrial DNA is an important issue of population and evolutionary genetics. Understanding the complex mechanisms by which mutations occur spontaneously in mtDNA of various mammalian organisms is an important goal of molecular biology.

Results: We have analyzed the somatic mutations spectrum of hepatic mtDNAs from OXYS and Wistar rats and have found that transitions constitute the most part of base substitutions in the mutation spectrum of both rat strains (94 % and 97 % of all mutations for OXYS and Wistar rats, respectively). Our results suggest the different mechanisms by which the mutations occur in the two studied rat strains.

INTRODUCTION

Mutation frequencies vary significantly along nucleotide sequences sometimes resulting in the formation of so called hotspots – the certain positions with high frequency of mutations. Mutation hotspots in DNA reflect the intrinsic properties of the mutation process, such as sequence specificity, that manifests itself at the level of interaction between mutagens and DNA, and the action of the repair/replication machineries. The nucleotide sequence context of mutational hotspots is a fingerprint of interactions between DNA and repair/replication/modification enzymes, and the analysis of the hotspot context provides evidence of such interactions. The hotspots might also reflect structural and functional features of the corresponding DNA sequences and provide the information about natural selection.

The most variable part of human and mouse mtDNA is a noncoding region (control region) which spans 1122 bases between the tRNA genes for proline and phenylalanine, and including D-loop region (Anderson *et al*., 1981). To present day there is no any investigation of generation and accumulation spontaneous somatic mutations in this region for rats. Therefore, it was of great interest to verify whether D-loop region is also a hypervariable segment (HVS) for rat's mtDNA.

Oxidative damage to cells has been regarded as a significant factor in carcinogenesis and aging. Therefore, special attention has been focused on the understanding the mechanism of oxidative damage. The strain of senescence-accelerated OXYS rats represents the valid model for studying ageing and neurodegenerative processes. The animals of this strain are characterized by a number of pathological manifestations similar to human geriatric disorders, e.g. early cataract and macular degeneration, senile osteoporosis, cardiomyopathy and behavioral disorders typical for aging animals and

humans (Salganik *et al*., 1994; Kolosova *et al*., 2003; Kolosova *et al*., 2004; Bobko *et al*., 2005). Mitochondrial dysfunctions increasing with age are supposed to be a causal factor for accelerated senescence in OXYS rats. Therefore, OXYS rat strain appears to be a useful model for studying ROS *in vivo* action on the genome and mutation process. In the present study we examined mitochondrial DNA somatic mutation spectra in the liver of OXYS and Wistar rats.

MATERIALS AND METHODS

Animals: 3-month-old male Wistar and OXYS rats obtained from the breeding experimental animal laboratory of Institute of Cytology and Genetics, RAS (Novosibirsk, Russia) were used in this study.

Purfication of total DNA: Total hepatic DNA was isolated from frozen tissues using the DNA extracted Medigen kit.

PCR/cloning procedure: Primers for D-loop were designed using GeneRuner program.

Forward primer (5′-atgaaattaatgtcccgatag-3′) was positioned from 15269 to 15291 nucleotides, and reverse primer (5′-ttaccaaccctgagaggtac-3′) was positioned from 312 to 292 nucleotides. The fragment of 1301 nucleotides contained $tRNA^{Thr}$, $tRNA^{Pro}$, D-loop, $tRNA^{Phe}$, 12S rRNA genes.

High fidelity "HF2" polymerase (Clontech) with proofreading activity (Advantage™ – HF2 PCR Kit) was used in PCR reactions. Cloning was performed with the pGem® T-easy Vector kit (Promega). T-vector was transformed into cells *E.coli* strain MRF. Individual clones were selected with blue-white selection. Cells were grown overnight in a shaking incubator at 37 °C in LB rich medium (Miller, 1972). Plasmid DNA was extracted by boiling technique and then sequenced using BigDye Terminator Premix v.3.1 (Applied Biosystems). We have sequenced 46 and 98 individual clones containing inserts of mtDNA fragments from Wistar and OXYS rats respectively.

On-line MultAline (http://prodes.toulouse.inra.fr/multalin/multalin.html) program were used to analyze the obtained sequences from different transformed *E. coli* clones. Using these data the mutation spectra for D-loop region of OXYS and Wistar rat's mtDNA were constructed. It was found that frequency of mutations in the control region of mtDNA for OXYS and Wistar rat strains constituted $1.7*10^{-3}$.

RESULTS AND DISCUSSION

The mutation frequencies for different types of base substitutions were not the same in the two studied rat strains: transitions constituted 94 % and 97 % of all mutations in the OXYS and Wistar mutation spectra, respectively (Table 1). This result is in a good agreement with previous observations for population polymorphism in humans (Vigilant *et al.*, 1991; Tamura, 2000; Malyarchuk *et al.*, 2002; Malyarchuk, Rogozin, 2004). We did not find any significant differences between the frequency of the different types of base substitutions between the OXYS and Wistar mutation spectra (Fisher exact test, $P = 0.81$).

We studied mechanisms of somatic mutations using the DNA context approach (Rogozin, Kolchanov, 1995). Strand slippage in repetitive sequences may result in base substitutions by the transient misalignment dislocation mechanism (Fig. 1). This model suggests that transient strand slippage in a monotonous run of nucleotides in the primer or template strand is followed by incorporation of the next correct nucleotide (Kunkel, Soni, 1988). Our earlier analysis of phylogenetically reconstructed mutation spectra (distributions of mutations along analyzed sequences) of the human mtDNA HVS I and II regions has suggested that the dislocation mutagenesis (Fig. 1) plays an important role for generating base substitutions in mtDNA (Malyarchuk *et al.*, 2002; Malyarchuk, Rogozin, 2004).

Additionally, it was found that next-nucleotide effects and dislocation mutagenesis may contribute to the formation of mtDNA mutations in patients with alterations of nucleoside metabolism (Nishigaki *et al.*, 2003). Therefore, the origin of mtDNA mutation hotspots may to a great extent depend on context properties of mtDNA. The dislocation mutagenesis operating in monotonous runs of nucleotides might play an important role for generating base substitutions in mitochondrial DNA and define context properties of mtDNA.

Table 1. Frequencies of transitions/transversions

	OXYS	Wistar		OXYS	Wistar
Transitions			Transversions		
A → G	0.27	0.34	A → C	0.	0.
T → C	0.34	0.43	T → G	0.03	0.03
G → A	0.14	0.	A → T	0.	0.
C → T	0.19	0.20	T → A	0.	0.
			G → T	0.	0.
			C → A	0.	0.
			G → C	0.03	0.
			C → G	0.	0.

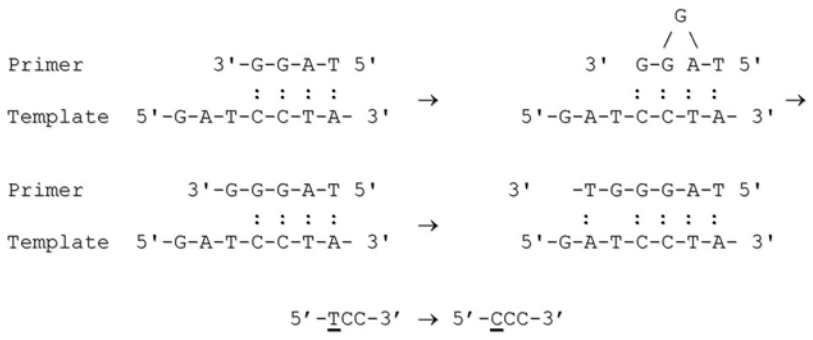


Figure 1. Dislocation mutagenesis. The primer strand dislocation during H-strand replication, three-nucleotide subsequences of the template strand are shown below schematic representation of the dislocation model.

The dislocation mutagenesis model was analyzed for the reconstructed mutation spectrum using a Monte-Carlo procedure (the KUNKEL program) (Rogozin, Kolchanov, 1995). Analysis of the Wistar spectrum revealed that many base substitutions (50 %) are consistent with the dislocation model. Furthermore, the primer strand dislocation model (Fig. 1) has a statistically significant support ($P(W \leq W_{random}) = 0.014$). The template strand dislocation mutagenesis also has a significant impact ($P(W \leq W_{random}) = 0.028$). A strong statistical support for the dislocation model suggested that a substantial fraction of somatic mutations (~50 %) are truly caused by dislocation mutagenesis.

No statistically significant support for the dislocation model was found in the OXYS spectrum ($P(W \leq W_{random}) = 0.68$). This result suggested that the dislocation mutagenesis does not play an important role for generating substitutions in OXYS rats. The observed differences in dislocation mutagenesis suggested that mechanisms of mutations in the two studied strains of rats are different.

CONCLUSION

The mutation frequency in investigation region of mtDNA for OXYS and Wistar rats is $1.7*10^{-3}$. Therefore this region may be considered as hypervariable segments.

Transition frequencies in mutation spectra of OXYS and Wistar mtDNAs constitutes more than 90 % of all substitutions and is in a good agreement with literature data on population polymorphism in human mtDNA. Analysis of the Wistar and OXYS rats spectrum revealed that mutations in the two strains occur by different mechanisms. In Wistar rats the most mutations are caused by dislocation mutagenesis.

ACKNOWLEDGEMENTS

The research was supported in part by grants from Russian Foundation for Basic Research (Nos 05-04-48779, 05-04-48483) and from Siberian Division of the Russian Academy of Science.

REFERENCES

Anderson S., Bankier A.T., Barrel B.G., de Bruijn M.H.L., Coulson A.R., Drouin J., Eperon I.C., Nierlich D.R., Roe B.A., Sanger F., Schreier P.H., Smith A.J.M., Staden R., Young I.G. (1981) Sequence and organization or the human mitochondrial genome. *Nature*, **290,** 457–467.

Bobko A.A., Sergeeva S.V., Bagryanskaya E.G., Markel A.L., Khramtsov V.V., Reznikov V.A., Kolosova N.G.(2005) 19F NMR measurements of NO production in hypertensive ISIAH and OXYS rats. *Biochem. Biophys. Res. Commun*., **330**(2), 367–370.

Kolosova N.G., Lebedev P.A., Aidagulova S.V., Morozkova T.C. (2003) OXYS Rats as a model of senile cataract. *Bull. Exp. Biol. Med.*, **136**, 415–419.

Kolosova N.G., Lebedev P.A., Dikalova A.E. (2004) *Bull. Exp. Biol. Med*., **137**, 249–251.

Kunkel T.A., Soni A. (1988) Mutagenesis by transient misalignment. *J. Biol. Chem.*, **263**, 14784–14789.

Malyarchuk B.A., Rogozin I.B. (2004) Mutagenesis by transient misalignment in the human mitochondrial DNA control region. *Ann. Hum. Genet*., **68**, 324–339.

Malyarchuk B.A., Rogozin I.B., Berikov V.B., Derenko M.V. (2002) Analysis of phylogenetically reconstructed mutational spectra in human mitochondrial DNA control region. *Hum. Genet.*, **111**, 46–53.

Miller J.H. Experiments in molecular genetics. Cold Spring Harbor Laboratory, 1972.

Nishigaki Y., Marti R., Copeland W.C., Hirano M. (2003) Site-specific somatic mitochondrial DNA point mutations in patients with thymidine phosphorylase deficiency. *J. Clin. Invest.*, **111**, 1913–1921.

Rogozin I.B., Kolchanov N.A. (1995) Computer system for analysis of nucleotide context role in mutagenesis (MutAn). *J. Cell. Biochem.*, Suppl. **21A**, 345.

Salganik R.I., Solovyova N.A., Dikalov S.I., Grishaeva O.N., Semenova L.A., Popovsky A.V. (1994) Inherited enhancement of hydroxyl radical generation and lipid peroxidation in the S strain rats results in DNA rearrangements degenerative diseases and premature aging. *Biochem. Biophys. Res. Commun*., **199**, 726–733.

Tamura K. (2000) On the estimation of the rate of nucleotide substitution for the control region of human mitochondrial DNA. *Gene*, **259**, 189–197.

Vigilant L., Stoneking M., Harpending H., Hawkes K., Wilson A.C. (1991) African populations and the evolution of mitochondrial DNA. *Science*, **253**, 1503–1507.

REFINEMENT OF PHYLOGENETIC SIGNAL IN MULTIPLE SEQUENCE ALIGNMENT: RESULTS OF SIMULATION STUDY

Rusin L.Y.*[*]*, Lyubetsky V.A.
Institute for Information Transmission Problems, RAS, Moscow, Russia
[*] Corresponding author: e-mail: roussine@yandex.ru

SUMMARY

Motivation: Disparate substitution rates within the different regions of homologous sequences and mutational saturation are well known to cause misalignment of sequences and to hamper accurate tree reconstruction. Therefore, there is a need in tools detecting and filtering out informational noise from the multiple alignment of sequence data; the tools will help to increase accuracy and resolution of phylogenetic analyses.

Results: We propose such a tool and tested its ability to improve the quality phylogenetic trees both on the biological COG data, and on the artificial data, where the ideal tree was known *a priory*. The key operation of the filtering is a removal of noisy columns. It was shown that the tool permits to reconstruct a tree closer to the "true" tree than is the tree reconstructed with data without removal. Procedure can be applied as a tool to pre-process multiple alignments and enhance phylogenetic inference.

INTRODUCTION

A common problem with large scale phylogenetic analyses is quality of primary sequence data. Many genomic applications require comparison of multiple phylogenies estimated from different families of orthologous genes in order to infer evolutionary events on a genomic scale. Prediction strength of this type of analysis in many respects will therefore depend upon reliability of individual reconstructions. Disparate substitution rates across regions of homologous sequences and mutational saturation are well known to result in elevated levels of homoplasy in the data and to overshadow available phylogenetic signal. The authors developed a procedure to detect and filter out informational noise from multiple alignment of protein sequence data, thus allowing one to considerably increase accuracy and resolution of phylogenetic analysis. In this work the procedure performance is studied with computer simulations.

MATERIALS AND METHODS

Phylogenetic software used at successive steps of the procedure was described in (Lyubetsky *et al.*, 2005) and includes originally developed programs that implement algorithms of computing the objective scoring function and constrained generation of random trees.

Simulations were conducted with the *evolver* program from PAML package (Yang, 1997). We have generated 1000 datasets, each consists of 40 amino acid sequences of length 300; the maximum-likelihood model parameters and branch lengths were obtained

from analysis of COG data, the same parameters were used in our previous studies (Lyubetsky *et al*., 2005).

Algorithm of the procedure was described in detail in (Lyubetsky *et al*., 2005). In essence, it uses a scoring function to rank columns of the alignment according to the consistency of the column's content with a *list of reliable clades* and gradually removes the least consistent ones until signal is refined to provide for better resolution of the tree. The list of reliable clades is basically the list of splits occurring in 70 % majority-rule consensus topology constructed after bootstrapping the intact alignment (i.e. before column removal). On each step of removing columns the g_1 statistic is estimated on current alignment (Hillis, Huelsenbeck, 1992) with the original algorithm of generating random trees strictly *compatible* with the list of reliable clades (Lyubetsky *et al*., 2005) and is used to determine the step, at which the procedure halts. The obtained alignment is considered optimal for tree reconstruction (definitive phylogenetic analysis).

SIMULATION RESULTS

Simulation studies were aimed at proving two statements: (1) removal of noisy columns permits to reconstruct a tree closer to the known tree than is the tree reconstructed with data without removal; (2) g_1 statistic estimated with the original algorithm of constrained random tree generation can be used to identify the refinement step, at which the procedure should be stopped.

The lists of reliable clades were quite different among the generated datasets, probably, due to unequal fraction of hypervariable sites retained in each of 500 replicates after bootstrapping the data. The datasets that produced well resolved consensus trees after bootstrapping were assumed to contain low amount of hypervariable sites and enough informative sites to produce a robust tree. Therefore, we used datasets (512 out of 1000), which produced consensus trees sufficiently unresolved to generate 100,000 constrained random topologies on their basis as test datasets to refine the signal.

Batch refinement of *in silico* generated datasets was continued for 10 steps. At each step, current alignment was analyzed to produce a phylogenetic tree and a g_1 score. Trees from successive steps were computed likelihoods against the *intact data* and compared using standard tests of phylogenies (approximately unbiased test, Kishino-Hasegawa test, Shimodaira-Hasegawa test).

In 100 % cases removing noisy columns permitted to reconstruct the tree, which is closer to the known tree used to simulate the data than is the tree obtained without refinement. In 91 % cases the tree with highest likelihood ("best" tree) was reconstructed at the step of the procedure, where the alignment produced the minimal (optimal) g_1 score, and in 53 % cases the difference in likelihood between the found "best" tree and the tree inferred with intact data was statistically significant. In 9 % cases the g_1 score continued to decrease beyond the step, at which the "best" tree is found, which might suggest that, although the signal related to poorly resolved branches of 70 %-consensus can be refined further, the columns needed to correctly reconstruct shallow parts of the whole tree (containing recent evolutionary events and, therefore, described by more variable regions) are already removed. Further studies will be conducted to develop measures of clade-specific noise removal. In the meantime, the described procedure can be used to refine alignments and improve phylogenetic inference with the advice to compare likelihoods of trees before and after column removal.

ACKNOWLEDGEMENTS

The authors are greatly indebted to O. Zverkov for valuable help.
The work was partly supported by grants RFBR Nos 05-04-49705 and ISTC 2766.

REFERENCES

Hillis D.M., Huelsenbeck J.P. (1992) Signal, noise, and reliability in molecular phylogenetic analyses. *J. of Heredity*, **83**, 189–195.

Lyubetsky V.A., Gorbunov K.Y., V'yugin V.V., Rusin, L.Y. (2005) Removing noise in multiple protein alignment. *Information processes*, **5**(5), 380–391.

Yang Z. (1997) PAML: a program for phylogenetic analysis by maximum likelihood version. *CABIOS*, **13**, 555–556.

GENETIC DIVERSITY AND PHYLOGENETIC RELATIONSHIPS IN GROUPS OF ASIAN GUARDIAN, SIBERIAN HUNTING AND EUROPEAN SHEPHERD DOG BREEDS

Ryabinina O.M.
Vavilov Institute of General Genetics, RAS, Moscow, Russia, e-mail: Olga_riabinina@newmail.ru

Key words: dog, mtDNA, phylogeny, genetic diversity

SUMMARY

Motivation: Analysis of Asian dog breeds has an important role in understanding phylogenetics of many dog breeds, because an East Asian origin of domestic dogs has been established.

Results: It was shown extremely close phylogenetic relationships in group of Asian guardian dog breeds and close relation to this group breed German Shepherd dog and group "Laika". One of Portuguese breeds (Serra da Estrella Mountain Dog) reveals more relation to Asian guardian dogs than to other Portuguese breeds. Breed groups Central Asian Sheepdog, Northern Caucasian Volkodav and Laika characterized by relatively high level of genetic diversity in contrast with Caucasian Ovtcharka.

INTRODUCTION

The origin of domestic dog and phylogenetics of dog breeds is very intriguing and complicated problem. Late years it was shown that dog has an origin from several wolf lines, most of all were distributed in East Asia (Savolainen *et al*., 2002). Modern dogs have mtDNA haplotypes belonging to five mtDNA haplogroups (A, B, C, D, E). Three of them (A, B, C) include > 95 % of studied sequences. Haplogroups B and C in contrast with haplogroup A have star-like form of networks with central nodes. Many authors have shown that there is no breed – specific haplogroups of mtDNA (Vila *et al*., Okumura *et al*., and others).

In order to evaluate more detailed picture of some breed phylogenetic relationships we have chosen several Asian guardian dogs and combined them to groups according to theirs geographical distribution and origin. As a result we have following breed groups: a) Turkish dogs (akbash and kangal), b) Central Asian Sheepdog (CAS), c) Northern Caucasian Volkodav (NCV), d) Caucasian Ovtcharka (CO). By the same approach we have chosen group "Laika".

For comparison we selected Portuguese breeds Portuguese Sheepdog (PS), Serra da Estrella Mountain Dog (SEMD), Azores Cattle Dog (ACD) and Dutch breed German Shepherd Dog (GSD). Different morphological type, employment, areas of origin and distribution characterize these breeds. By these reasons we analyzed them each as separate group.

MATERIAL AND METHODS

Blood and hairs from “CAS”, “NCV”, “CO”, “GSD” and “Laika” were collected during veterinary practice and at canine moves in Caucasian and Moscow regions. Left variable segment of mtDNA was amplified with primers L15910 and H16498 (DeSalle *et al*., 1993). Sequencing of gel-purified PCR products was performed using BigDye Terminator Kit v. 3.1 on ABI-Prism 3100-Avant instrument (Applied Biosystems, USA).

Nucleotide sequences and frequencies of haplotypes in breeds PS, SEMD, ACD, “Turkish dogs”, some German shepherds and “Laika” were taken from published papers (Savolainen *et al*., 2002, van Ash *et al*., 2005) and GenBank.

Nucleotide diversity, gene diversity, Fst, Tamura-Nei genetic distances and AMOVA were calculated using Arlequin v.3.0 (Excoffier *et al*., 2005). Haplotype network was constructed using TCS v.1.21 (Clement *et al*., 2000).

RESULTS AND DISCUSSION

In group “CAS” we identified 14 haplotypes belonging to A (9), B (2), C (1), W/E (2) clades, in “CV” – 12 belonging to clades A (7), B (2), C (2), D (1), in “CO” – 5 haplotypes, belongs to clades A (3) and B (2). We identified 10 novel haplotypes, 6 of them belongs to clade A, 1 – to clade B, 1 – to clade C and 2 of them belong to clade W/E. It is interesting, that haplotypes from clade E closely related to wolf clade W were found before only in Japanese and Korean breeds.

As was previous shown (Savolainen *et al*., 2002) haplogroup D has maximum diversity and high frequencies in Scandinavia and occasionally observed in Spain, Portugal and Turkey. Occurrence of early-identified haplotype D6 (RH20) in “CV” and D1-4 (RH25) in “Laika” may indicate a gene flow from Scandinavian breeds to breeds of these regions through Europe and Asia Minor.

CAS, CV and Turkish breeds share 3 common haplotypes, furthermore one haplotype RH1 (A11-14, 54) that occurs with high frequencies in “CAS”, “CV”, “CO”, “Laika” and “Turkish breeds” is absent in Portuguese breeds. As we show in Fig. 1 haplogroup D may originate from this haplotype.

Pairwise differences between groups measured as Fst are zero and therefore indicate absence of differentiation between “CAS”, “NCV”, “CO”, “Turkish dogs”, “Laika” and “GSD”. One of Portuguese breed - SEMD - shows Fst about 0,075 with these groups in contrast with 0,25 and 0,30 with other Portuguese breeds ACD and PS respectively. This fact reflects high level of differentiation between Portuguese breeds and less differentiation with groups specified above.

For phylogeny inference we calculated Tamura-Nei genetic distances taking into account different transition/transversion rates, differences in base frequencies and transition rates between purines and between pyrimidines. As shown in Fig. 2 “Laika”, “CO” and “GSD” slightly stand out from pointed out nondifferentiated group.

AMOVA indicates the following: 1) combining all breeds in one group 16,36 % was among breeds component and 83,64 % – within breed, 2) combining “CAS”, “NCV” and “Turkish dog” in one group, “CO” and “PS” in another, and others breeds each in separate group the level of within-groups component was the minimal and among-group’s was maximal values in comparison with other possible combinations based on genetic distances. 3) In case of groups of breeds were combined according to geographical regions of theirs distribution, among group component was always negative and nonsignificant.

Gene diversity was high in groups “CAS”, “NCV”, “Laika” (0,83–0,88) middle in “GSD” and “Turkish dogs” (0,70–0,76) and low in “PS”, “ACD” and “CO” (0,48; 0,64; 0,65). Low gene diversity in breeds subjected to cultural breeding may be due by prevalent inbreeding.

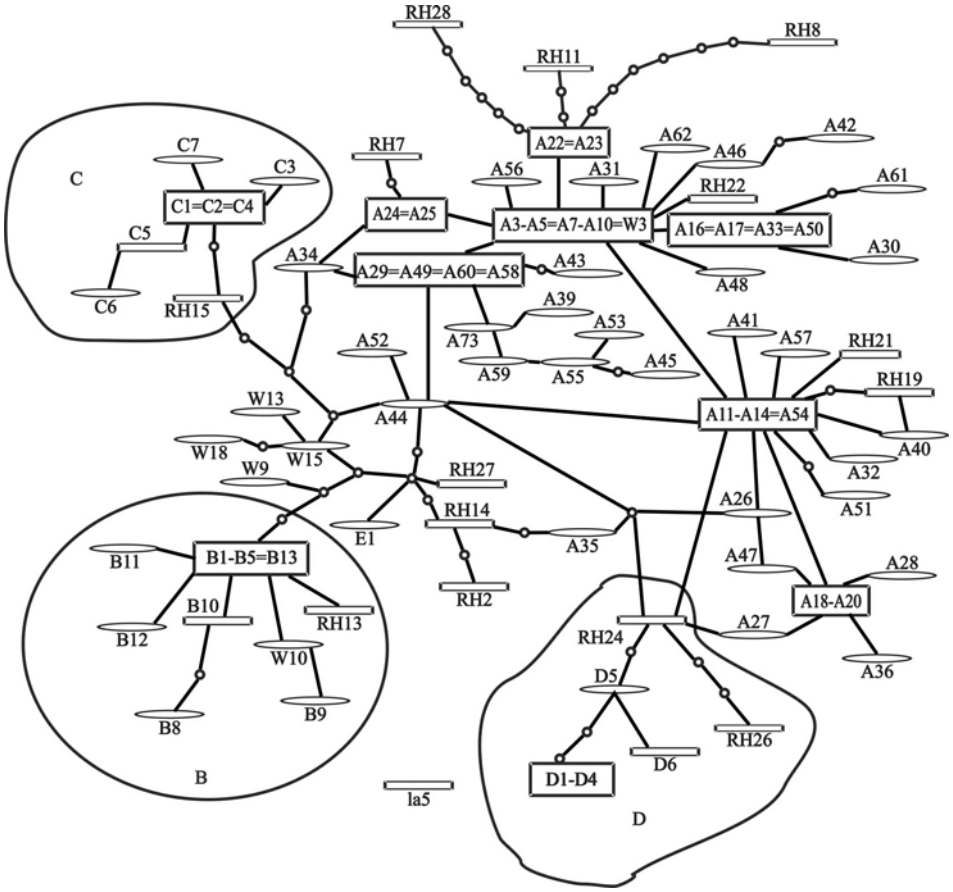


Figure 1. Haplotype network of dog mtDNA D-loop haplotypes. Haplotypes typed in rectangles were detected in “CAS”, “CV”, “CO” and “Laika”. Haplogroups B, C, D are whirlpooled.

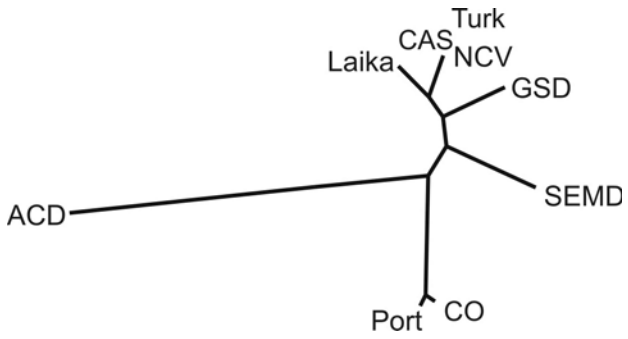


Figure 2. UPGMA tree based on Tamura-Nei distances.

Our results indicate major role of gene flow between regions during forming of dog breeds. We have shown a possibility of Asian origin of European breed German Shepherd Dog and high involvement “Asian” guardian dog lines in genesis of Portuguese breed Serra da Estrella Mountain Dog.

Genetic data presented in this work have historical and archeological confirmations.

ACKNOWLEDGEMENTS

I would like to acknowledge corresponding member of RAS Prof. Ilya A. Zakharov for helpful comments on the work. Thanks to Olga Krasnovskaya for providing hairs from "Northern Caucasian Volkodavs" and Dr. Vladimir K. Bojenko for access to ABI-Prism 3100-Avant.

This work was supported in part by the leading scientific schools grant NSh-10122.2006.4 of President of Russian Federation.

REFERENCES

Clement M. *et al*. (2000) TCS: a computer program to estimate gene genealogies. *Mol. Ecol.*, **9**, 1657–1660.

DeSalle R. *et al.* (1993) Isolation and characterization of animal mitochondrial DNA. *Methods Enzymol.*, **224**, 176–204.

Excoffier L. *et al*. (2005) Arlequin ver. 3.0: An integrated software package for population genetics data analysis. *Evolutionary Bioinformatics Online*, **1**, 47–50.

Savolainen P. *et al.* (2002) Genetic evidence for an East Asian origin of domestic dogs. *Science*, **298**, 1610–1613.

van Asch B. *et al*. (2005) MtDNA diversity among four Portuguese autochthonous dog breeds: a fine-scale characterization. *BMC Genetics*, **6**, 37–44.

ANALYSIS OF EGFR GENE MUTATIONS WHICH HAVE A RESPONSE TO QUINAZOLIN INHIBITORS

***Sabitha K.*[1, 2]*, Kaiser J.*[*1, 2]**

[1] Genetics Department, Bhagawan Mahavir Hospital and Research Center, 10-1-1, Masab tank, Hyderabad-500 004 A.P, India; [2] Indo American Cancer Institute and Research Center, Banjara Hills, Hyderabad, 500034, A.P., India

[*] Corresponding author: e-mail < Kaiser.jamil@gmail.com >

Key words: mutations, affinity, homology modeling, *in silco*, cheminformatics, EGFR, quinazolin, ligands

Abbreviations: EGFR-Epidermal growth factor receptor; WT-wild type; u.c. - unit cell.

Motivation: EGF receptor and its ligands are involved in over 70 % of all cancers. Mutations of the EGFR gene have been identified in specimens from patients with non-small lung cancer who have a response to quinazolin inhibitors such as CI1033, Gefitinib, and Erlotinib. Substitution mutations G719S and L858R were reported by Guillermo Paez *et al*. (2004). Both types of mutation increase the sensitivity of the tumor to quinazoline inhibitors of EGFR. This prompted us to take up this study on mutation of EGFR gene and its sensitivity to quinazolin inhibitors using cheminformatics.

Methodology: The homology modeling was carried out with the Insight II Modeller module. The interaction energy of the quinazolin inhibitors with each individual amino acid in the active site of EGFR was calculated by the advanced program Affinity.

Results and Validation: The crystal structure of Syk Tyrosine Kinase Domain was used as template in the modeling of EGFR Tyrosine Kinase Domain (mutated sequences). Ramachandran plot statistics indicated that 91 % G719S, 89 % L858R of the main-chain dihedral angels were found in the most favorable region. The surface area of mutant EGFR G719S, L858R and WT EGFR active sites were 657.73 $Å^2$/u.c., 753.76 $Å^2$/u.c. and 466.36 $Å^2$/u.c. respectively. The overall binding affinity of the ligands was slightly affected, but surprisingly the mutated serine residues were found to have better interaction with C11033 and Gefitinib but not with Erlotinib.

INTRODUCTION

EGFR expression is known in normal epithelial tissues like skin, hair follicle, and gastrointestinal tract, and is important for normal cell functions. Over expression or co expression of the EGFR and its ligands enhance the activity of receptor which is the hallmark of many human carcinomas (Yarden, 2001). In many types of tumors, including lung, breast, prostate, ovary, gastrointestinal tract and brain, the EGFR receptor is expressed approximately 100-times more than those expressed in normal cells. Over expression of at least two of these receptors, the EGF receptor and the closely related ErbB2 has been associated with a more aggressive clinical behavior (Alroy, Yarden, 1997). Indeed, expression of high levels of these two receptors in nonmalignant cell lines leads to a transformed phenotype. EGFR tyrosine kinase is therefore a potential target for various cancer therapies (Yaish *et al*., 1988; Levitzki *et al*., 1999; Baselga, 2002).

METHODS

Databases, Equipment, and Software

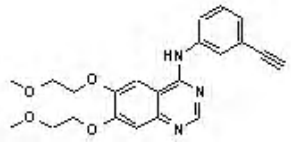
The databases used in the study include the EMBL, Genbank, Swissprot, PDB and PROSITE. Protein homology -was done on Silicon Graphics Octane and O2 workstations (Silicon Graphics Inc., Mountain View, CA). Insight II 98.0 software was used for the protein homology modeling and Procheck for structure validation. The software utilized in the computational analysis was GRID Chemical structures were imported from ISIS-BASE database or drawn in ISIS-Draw (MDL information Systems Inc., San Leandro, CA).

The crystal structure of Syk Tyrosine Kinase Domain was used as template in the modeling of Epidermal Growth Factor Receptor (EGFR) Tyrosine Kinase Domain (mutated sequences). The sequence identity between Syk and EGFR is 64 % which makes Syk Tyrosine Kinase Domain a good template for the modeling of EGFR. The modeling was carried out with the Insight II Modeller module.

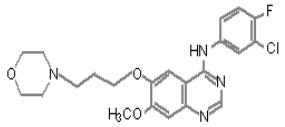
Amino acid sequence of EGFR was obtained from SWISS-PROT database, accession no P00533. Sequence homology searches were carried out using BLAST algorithm against Protein data bank. Target-template alignment was created using ALIGN2D command of Modeller which implements a global dynamic algorithm with a variable gap penalty. The alignment is done to the mutated protein. Modeling was carried out selecting 1XBC (Syk Tyrosine Kinase Domain) as the primary template. Both share a sequence similarity 64 % to enhance the quality of protein in terms of the structurally conserved regions and structurally variable regions. We selected 1FPU.pdb, 1M52.pdb, 1U54.pdb as secondary template. Sequence alignment of all the sequences was performed using CLUSTAL W.

MOLECULAR 3D STRUCTURE BUILDING

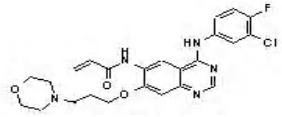
Ligands considered for docking were built by model builder in Cerius2 and minimized using conjugate gradient algorithm with a gradient convergence value of 0.01 Kcal/mol Å. Partial atomic charges were calculated using the Gasteiger-Hückel method (Gasteiger, Marsili, 1980) Further geometry optimization was carried out for each compound with the semi-empirical method using the AM1 Hamiltonian in MOPAC 6 in Cerius2. (Dewar *et al*., 1985). Initially, a constrained minimization for 100 cycles was performed in which 3 rings were defined as an aggregate to constrain their conformation to avoid false minima. The constraints were then removed, and the structure was subjected to 1000 cycles of minimization or till the gradient converged to 0.001 Kcal/mol Å.



Erlotinib



Gefitinib



CI-1033

MOLECULAR DOCKING (AFFINITY)

The crystal structure of human, EGFR tyrosine kinase domain (1M17) (Stamos *et al*., 2002) with the inhibitor erlotinib was obtained from protein databank. H-atoms were added to the proteins while keeping all the residues in charged form. The complex was subjected to minimization using steepest descent (1000 iterations) followed by conjugate gradient (1000 iterations) algorithms with CVFF force field. Implicit salvation parameters were considered. Amino acid residues in the region of 10 Å radius around the bound

ligand were selected and allowed to be flexible while keeping rest of the protein rigid. For each molecule various conformations were generated and the electrostatic and steric energies were calculated. The most favorable conformation of each ligand in the EGFRWT and mutant EGFR complex is chosen for further analysis. The structural information from the theoretical modeled complex can help us to understand the catalytic mechanism of enzyme.

RESULTS AND DISCUSSION

Mutations of the EGFR gene have been identified in specimens from patients with non-small lung cancer who have a response to quinazolin inhibitors such as CI1033, Gefitinib, and Erlotinib. The substitution mutations G719S and L858R are located in the GXGXXG motif of the nucleotide triphosphate binding domain or P-loop and adjacent to the highly conserved DFG motif in the activation loop. This mutation mediates oncogenic effects by altering downstream signaling and antiapoptotic mechanisms. This glycine residue was mutated to Serine residue by using in silico technique. The 3-D structures of the mutant EGFR G719S and L858R were built by using crystal structure coordinates of SYK (1XBC.PDB) as template using the program- Modellar. This program is completely automated and is capable of generating energy minimized protein models by satisfying restraints on bond distances and dihedral angels. Each model was subjected to various cycles of Modeller and the best possible model was selected. Model evaluation was conducted using ENERGY command of Modeller, program PROCHECK. Ramachandran plot statistics indicated that 91 % G719S, 89 % L858R of the main-chain dihedral angels were found in the most favorable region.

The homology model was then used for active site docking of compounds with known orientation toward the Erlotinib, the selected conformers were used as templates in a similarity analysis. The remaining compounds were then docked into the homology model and the resulting conformers were evaluated in the similarity analysis (based on grid interaction energies) to choose conformers similar to the templates. This approach enabled a conformer selection that constrained the conformational space to the active site of the homology model. The active site of the mutant EGFR structures was analysed and compared to EGFR crystal structure (1M17.pdb) with respect to structural difference. The active site surface area of mutant EGFR G719S and L858R are 657.73 $Å^2$/u.c. and 753.76 $Å^2$/u.c. and for WT EGFR active site it is 466.36 $Å^2$/u.c.

The interaction energy of the quinazolin inhibitors with each individual amino acid in the active site of EGFR was calculated by the advanced program Affinity. Residues LEU_A694/23, GLY_A695/24, VAL_A702/31, ALA_A719/48, LEU_A764/93, THR_A766/95, GLN_A767/96, LEU_A768/97, MET_A769/98, GLY_A772/101, THR_A830/159, ASP_A831/160, GLY_A833/162 (crystal structure residue numbering/ homology modeled structure residue numbering) where seen to have significant interactions with the ligand and are important for binding. Here we observed one interesting thing; the mutated ligands in G719S did not interact with MET_A742/71. Also EGFR crystal structure LEU_834/163 did not interact with the ligands; but in homology models all the three ligands were found to interact. The overall binding affinity of the ligand molecule was slightly affected, but surprisingly the mutated serine residue was found to have better interaction with ligand molecules except with Erlotinib. Interaction energies (electrostatic and steric) of EGFR WT Gly719 Etotal were -0.161 Kcal mol^{-} (CI1033), -0.877 Kcal mol^{-} (Gefitinib), -0.951 Kcal mol^{-} (Erlotinib), and for EGFR mutant Ser24 Etotal were -1.314 Kcal mol^{-} (CI1033), -1.504 Kcal mol^{-} (Gefitinib),-0.786 Kcal mol^{-} (Erlotinib). In crystal structure Lue858 did not interact with ligands but in Arginine mutation it interacted with ligands and gave good affinity. So Gefitinib and CI1033 showed good sensitivity in mutated EGFR.

This study determines the importance of mutations in EGFR gene pharmacogenomics, and it is useful information for both the patient and the clinician. This study also helps us to identify the responses and no responses to the drugs which may lead towards personalized medicine.

REFERENCES

Alroy I., Yarden Y. (1997) The ErbB signaling network in embryogenesis and oncogenesis: signal diversification through combinatorial ligand-receptor interactions. *FEBS Lett*., **410**, 83–86.

Baselga J. (2002) Why the epidermal growth factor receptor? The rationale for cancer therapy. *Oncologist*, **7**, 2–8.

Dewar M.J.S., Zoebisch E.G., Healy E.F., Stewart J.J.P. (1985) AM1: A new general purpose quantum mechanical molecular model. *J. Am. Chem. Soc*., **107**, 3902–3909.

Gasteiger J., Marsili M. (1980) Iterative partial equalization of orbital electronegativity. *A rapid access to atomic charges*., **36**, 3291–3228.

Levitzki A. (1999) Protein tyrosine kinase inhibitors as novel therapeutic agents. *Pharmacol. Ther*, **82**, 231–239.

Paez G.J., Janne P.A., Lee J.C., Tracy S., Greulich H., Gabriel S., Herman P., Kaye F.J., Lindeman N., Boggon T.J., Naoki K., Sasaki H., Fujii Y., Eck M.J., Sellers W.R., Johnson B.E. (2004) Matthew eyerson. EGFR Mutations in Lung Cancer: Correlation with Clinical Response to Gefitinib Therapy. *Science*, **304**, 1497–1500.

Stamos J., Sliwkowski M.X., Eigenbrot C. (2002) Structure of the epidermal growth factor receptor kinase domain alone and in complex with a 4-anilinoquinazoline inhibitor. *J. Biol. Chem*., **277**, 46265–46272.

Yarden Y. (2001) The EGFR family and its ligands in human cancer. Signalling mechanisms and therapeutic opportunities. *Eur. J. Cancer*, **37**(Suppl. 4), S3–8.

Yaish P., Gazit A., Gilon C., Levitzki A. (1988) Blocking of EGFdependent cell proliferation by EGF receptor kinase inhibitors. *Science*, **242**, 933–935.

TRANSPOSON-FREE REGIONS IN MAMMALIAN GENOMES

Simons C., Pheasant M., Makunin I.V., Mattick J.S.*
ARC Special Research Centre for Functional and Applied Genomics, Institute for Molecular Bioscience, University of Queensland, Brisbane QLD 4072, Australia
* Corresponding author: e-mail: j.mattick@imb.uq.edu.au

Key words: human genome, gene regulation, development, noncoding

SUMMARY

Motivation: Despite harbouring more than 3 million transposon insertions, the human genome has some very long regions without transposons. We performed a comprehensive analysis of extended transposon-free regions in the human and mouse genomes.

Results: We identified almost one thousand transposon-free regions (TFRs) over 10 kb in length in both the human and mouse genomes. The majority of human TFRs correlate with orthologous TFRs in the mouse, and many also overlap with orthologous TFRs in the opossum, despite the fact that in primates, rodents and marsupials most transposons are lineage specific. Over 90 % of the bases within TFRs are noncoding, much of which is not highly conserved. TFRs associate with genes involved in development and presumably represent extended regulatory regions.

Availability: FASTA files with description and sequence of each TFR are available at http://jsm-group.imb.uq.edu.au/tfr/.

INTRODUCTION

The human genome has ~3.2 million recognizable transposons (SINEs, LINEs, DNA- and LTR-transposons) which comprise approximately half of human genome. The average distance between transposon insertions is 476 bp. Nevertheless, there are some long regions lacking any recognizable transposons, first identified around HOX gene clusters that are prominent developmental regulators. We therefore undertook a structured study to catalog and analyze very long transposon-free regions (TFRs) in the human and mouse genomes.

METHODS

M&M are described in Simons *et al*., 2006.

RESULTS AND DISCUSSION

We extracted all regions from human genome over 10 kb without any annotated transposons and then filtered out regions containing more than 20 % of non-transposon repeats, self-hits or mitochondrial inserts. The filtering allowed us to exclude putative cases of recent DNA expansion. The resulting set contains 860 transposon free regions

(TFRs) over 10 kb in length, covering over 12 Mb of the human genome, with the longest TFR over 81 kb in length. Table 1 lists the ten longest human TFRs. Similar procedures applied to mouse genome yield 993 TFRs covering over 13 Mb. TFRs in our final set contain primarily complex unique DNA sequences (less than 4 % repeats on average).

We estimated probability of observing this number of TFRs assuming either random or GC-dependent transposon distribution within the genome. In both models, the observed number of TFRs was significantly higher than expected (for details see Simons *et al.*, 2006). In addition, 47 % of human TFRs overlap with mouse TFRs. We noted that in the mouse genome TFR length is very often compromised by assembly gaps. Accounting for this we found that 85 % of human TFRs ≥ 10 kb overlap a mouse TFRs ≥ 5 kb. Moreover, 52 % of human TFRs ≥ 10 kb overlap TFRs ≥ 5 kb in the draft opossum genome. Considering that majority of transposon insertions in these organisms are independent we concluded that TFRs have remained resistant to insertions over a long period of evolutionary time, at least since the divergence of marsupials and placental mammals. At least some TFRs are conserved in chickens. Furthermore, the presence of TFRs in paralogous regions, such as all four HOX gene clusters and five out of six IRX genes suggests that these regions may have been refractory to transposons before the first whole genome duplication in the vertebrate lineage, estimated to have occurred more than 600 million years ago (Vandepoele *et al.*, 2004).

The majority of human TFRs (85 %) overlap one or more annotated genes, usually (80 %) including the 5′- and/or 3′-UTR. On average, 15 % of bases covered by TFRs are annotated as exonic, and 8 % as protein-coding. Most of the sequence within TFRs is intronic (43 %) or intergenic (42 %). Note that an average distance between annotated transposons within human genes is 525 bp, and the existence of TFRs could not be explained by lower transposon density within genes.

Genes with GO annotation 'transcription factor activity' show 4.4 fold enrichment ($P < 10^{-66}$) in association with TFRs (167 transcription factors from the 641 genes with an assigned GO annotation). The enrichment is higher for longer TFRs (≥ 15 kb). Genes associated with TFRs include members of IRX family of homeobox transcription factors (all but IRX6 are associated with TFRs between 10 and 22 kb), HOX, PAX, FOX (forkhead box), SOX (HMG1/2 box), LHX (LIM / homeobox), POU, SIX, TBX and ZIC gene families, all of which play key roles in development. Interestingly, 29 human miRNAs (out of 321) are contained within TFRs, representing a 23 fold enrichment.

TFRs are enriched with recently described ultra-conserved (uc) elements (Bejerano e*t al.*, 2004), with 87 out of 481 uc elements located within TFRs. We asked whether existence of TFRs could be explained by conservation of underlying DNA sequence. It turned out that 29 % of bases within TFRs correspond to conserved elements (sequences under negative selection) within the human genome as determined by A. Siepel and co-authors (2005) with the remaining 71 % of bases having little evidence for selection on primary DNA sequence. Moreover, 277 human TFRs contain more than 10 kb of 'non-conserved' bases, and even if we assume that transposon integration is allowed only within the "non-conserved" fraction of TFRs, the chance observing 277 TFRs by chance is still close to zero. This suggests that existence of TFRs cannot be explained by conservation of primary DNA sequence.

The GC content of the human TFRs exhibits a broadly bimodal distribution with a minimum between the peaks at 57 % GC content (see Supplementary materials from Simons *et al.*, 2006), and mouse TFRs show similar distribution. We arbitrarily divided the human data set based on GC content and analysed two sets separately. A "high GC" set of 245 human TFRs (GC content ≥ 57 %) shows no enrichment in association with transcriptional factors but high proportion of these TFRs are located close to telomeric regions: 50 % are located within 2 Mb of annotated chromosome ends and 71 % within 10 Mb (compared to 2 % and 14 % respectively of "lower GC" TFRs). A "lower GC" set of 615 TFRs have a 4.9 fold enrichment for genes annotated 'transcription factor activity'

($P < 10^{-50}$) as well as strong enrichment for uc elements. These observations suggest that TFRs may be occur for more than one reason.

We have described and analysed transposon-free regions in human and mouse genomes. The TFRs associate with developmentally important genes and presumably represent extended regulatory regions. Although TFRs are enriched in conserved sequences, their existence could not be explained by constrains on primary sequences. TFRs contain various complex DNA sequences, which suggest that those regions are under selection against transposon insertions rather than just being refractory to transposon integration. The existence of long regulatory regions (10 kb and more) resistant to transposon insertion but tolerant to significant amount of nucleotide substitutions is hard to reconcile with orthodox models of genome (gene) regulation. We suggest two non-exclusive roles for these regions. TFRs might be regions important for chromatin modification, and transposon insertions could change the epigenetic pattern of DNA and histone modification in these regions. Alternatively, TFRs may harbor regions that express functionally important non-coding RNAs that may be intoerant of insertions but not to substitutions. Whatever the molecular mechanisms involved, the presence of TFRs identifies large, presumably regulatory, regions that are important to mammalian ontogeny but that would otherwise be difficult to detect using traditional computational approaches based on primary sequence conservation.

Table 1. List of the ten longest human Transposon-Free Regions

TFR ID	Genomic Position	Length (kb)	% GC	Overlapping Genes	% exonic bases
chr7.119	chr7:26938206	81,2	52	HOXA4-11	17
chr5.354	chr5:92928424	57,7	46	NR2F1	14
chr2.598	chr2:176777785	53,9	52	HOXD8-13	18
chr11.129	chr11:31759920	46,9	49	PAX6	5
chr17.162	chr17:43994277	46,2	52	HOXB4-6	19
chr13.226	chr13:99405192	43,7	52	ZIC2, ZIC5	12
chr5.305	chr5:87986303	40,1	44	mir-9-2	0
chr2.316	chr2:104910522	39,3	49	POU3F3	4
chr15.252	chr15:94658960	38,0	48	NR2F2	8
chr7.319	chr7:96264636	37,6	49	DLX5	4

ACKNOWLEDGEMENTS

This research was supported by the Australian Research Council and the Queensland State Government.

REFERENCES

Bejerano G., Pheasant M., Makunin I., Stephen S., Kent W.J., Mattick J.S., Haussler D. (2004) Ultraconserved elements in the human genome. *Science*, **304**, 1321–1325.

Siepel A. *et al*. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*., **15**, 1034–1050.

Simons C., Pheasant M., Makunin I.V. Mattick J.S. (2006) Transposon-free regions in mammalian genomes. *Genome Res*., **16**, 164–172.

Vandepoele K., De Vos W., Taylor J.S., Meyer A., Van de Peer Y. (2004) Major events in the genome evolution of vertebrates: paranome age and size differ considerably between ray-finned fishes and land vertebrates. *Proc. Natl. Acad. Sci. USA*, **101**, 1638–1643.

EVOLUTIONAL AND FUNCTIONAL ANALYSIS OF T-BOX REGULON IN BACTERIA: IDENTIFICATION OF NEW GENES INVOLVED IN AMINO ACID METABOLISM

***Vitreschak A.G.*[*1], *Lyubetsky V.A.*[1], *Gelfand M.S.*[1]**
[1] Institute for Information Transmission Problems, RAS, Moscow, 127994, Russia,
[*] Corresponding author: l_veter@mail.ru

Key words: genome analysis, amino acid biosynthesis and transport, T-box, bacteria

SUMMARY

Motivation: T-box antitermiantion is the most distributed mechanism of regulation of various amino acids in Gram-positive bacteria. Identification of the T-box regulon and a metabolic analysis of amino acid biosynthesis and transport is one of problems of comparative genetics, genomics and molecular biology.

Results: Search for T-box elements and analysis of operon structures identified a large number of new candidate T-box regulated genes, mostly transporters, in Gram-positive bacteria. We assign amino acid specificity for a large number of candidate transporters as well as for other new amino acid related genes.

Availability: The program is available by request to the author.

INTRODUCTION

Computer comparative analysis is a powerful method of prediction of the RNA secondary structure. It has been used for prediction of both regulatory and structural RNAs. A somewhat different approach is to predict gene regulation by analysis of RNA patterns. We have used it to analyze the T-box regulatory elements in Gram positive bacteria. It is experimentally known a number of T-box elements in some Gram positive bacteria: *Bacilllus subtilis*, *Bacillus stearothermophilus*, *Lactococcus lactis* and *Staphylococcus aureus* and some others (Grundy *et al*., 1994; and others). Genes is known to be regulated by T-boxes encode in most cases aminoacyl-tRNA synthetases, amino acid biosynthetic operons and some amino acid transporters. The T-box regulatory element consists of the alternative RNA secondary structures (the terminator and antiterminator conformations) and a number of conservative sequences boxes. The uncharged amino acid-tRNA is the inducer of transcription. At low concentration of regulatory amino acid in medium it binds to the RNA structure (interacts with T-box and anti-anticodon site) and promotes formation of the antiterminator. In contrast, at high concentration of regulatory amino acid a terminator conformation forms that leads to premature termination of transcription.

DATA AND METHODS

Using the set of known T-box sites, we constructed the pattern of the T-box RNA element and scanned available genomic sequences using the RNA-PATTERN program

(Vitreschak *et al*., 2001) and another program, developed for these purposes (Leontiev, Lyubetsky, 2006). The input RNA pattern described the RNA secondary structure and the sequence consensus motifs. The RNA secondary structure was described as a set of the following parameters: the number of helices, the length of each helix, the loop lengths and the description of the topology of helix pairs.

RESULTS AND DISCUSSION

We found about 800 T-boxes in 90 bacterial genomes. T-boxes are widely distributed in Gram-positive bacteria (Firmicutes, Actinobacteria). Moreover, several T-boxes were found in some Gram-negative bacteria (δ-proteobacteria) and other groups (Dienococcales\Thermales, Chloroflexi, Dictyoglomi).

Comparison of sets of T-box-regulated genes in analysed genomes shows, that most genes is constituted by aminoacyl-tRNA synthetase genes. Two other groups of T-box regulated genes consist of amino acid biosynthetic genes and genes with unknown function. Distribution of T-boxes involved in regulation of aminoacyl-tRNA synthetase and amino acid biosynthetic genes by T-box antitermination is shown in Table 1.

Aminoacyl tRNA synthetase genes *ileS*, *valS*, *leuS*, *serS*, *thrS*, *pheST*, *alaS*, *asp*(*asn*)*S*, *glyS*(*QS*) are regulated by T-box antitermination in most Firmicutes and some other phylogenetic groups, whereas *metS*, *proS*, *CyS*, *hisS*, *argS*, *lysS* are regulated only in distinct groups/bacteria.

T-box antitermination mechanism is also involved in regulation of various amino acid biosynthetic genes. *trp* and *ilv(leu)* operons are found to be regulated in most Firmicutes as well as in some other groups. Other amino acid biosynthetic genes are regulated only in distinct groups/bacteria (Table 1). The conservation of the T-box antitermination in distinct groups can be explained by a variability of regulatory mechanisms. In particular, the methionine metabolism in Gram-positive bacteria was known to be controlled by five different mechanisms: S-box, T-box, metK-box regulation (acting on the level of premature termination of transcription/inhibition of translation initiation) and two other mechanisms acting on the DNA level (Met-box and MetJ-box) (Rodionov *et al*., 2004). In another case, regulation of genes of the aromatic amino acid biosynthesis pathway in Gram-positive bacteria is shown to be quite labile and involves at least four regulatory systems, two at the RNA level involving competition of alternative RNA secondary structures for transcription and/or translation regulation and two at the DNA level (Panina *et al*., 2003).

Positional analysis of T-boxes led to the identification of a large number of new candidate amino acid transporters (Table 2).

We predicted the amino acid specificity of possible transporters analyzing the T-box regulatory "specifier codon" (a T-box regulatory site involved in the interaction with the anti-codon site of the uncharged tRNA). The regulatory codon of the T-box RNA element is known to be located in the fixed internal loop of the specifier hairpin. We verified the amino acid specificity of all predicted T-boxes was by sequence and structural alignment (multAl, Mironov, unpublished) and construction of phylogenetic trees (In most cases, T-boxes with the same specificity located in the same branch of the T-box phylogenetic tree).

The predicted tyrosine specific transporter *yheL* (Na+/H+ antiporter) is found to be regulated by the (TYR)T-box antitermination in some Bacillales and Lactobacillales. A phylogenetic analysis showed that YheL form a separate branch on the NhaC superfamily phylogenetic tree. This family also includes lysine transporters LysW, methionine transporters MetT and malate/lactate antiporter MleN.

Table 1. Regulation of aminoacyl-tRNA synthetases and amino acid biosynthetic operons in Gram-positive bacteria

Amynoacyl-tRNA synthetases	
Aromatic a/a TRP, PHE, TYR	Most FIRMICUTES, *Atopobium minutum*
Branched chain a/a ILE, LEU,VAL	Most FIRMICUTES, Actinobacteria(ileS), Dienococcales\ Thermales(ileS, valS), Chloroflexi(ileS), *Thermomicrobium roseum*(leuS)
methionine	Bacillales, Clostridiales, *Thermoanaerobacter tengcongensis*
proline	Some Bacillales, Clostridiales,
cysteine	Bacillales, some Lactobacillales, Clostridiales, Thermoanaerobacteriales
histidine	Bacillales, Lactobacillales(exept streptococcus spp.), some Clostridiales, *Thermoanaerobacter tengcongensis*
arginine	Bacillales, Lactobacillales (exept streptococcus spp.), Clostridiales,
threonine	Bacillales, Lactobacillales, Clostridiales, Dictyoglomi, Thermomicrobium roseum
serine	Most FIRMICUTES
alanine	Bacillales, Lactobacillales, Clostridiales
ASP, ASN	Most FIRMICUTES (exept streptococcus spp., Mycoplasmatales, Entomoplasmatales)
glycine	Most FIRMICUTES, Dienococcales\ Thermales
lysine	*Bacillus cereus*, *Clostridium thermocellum*
Amino acid biosynthetic genes	
Aromatic a/a TRP, PHE, TYR	Most FIRMICUTES, Chloroflexi and Dictyoglomi (trp operon), some FIRMICUTES (aro genes, pheA, pah)
Branched chain a/a ILE, LEU,VAL	Bacillales, Clostridiales, *Syntrophomonas wolfei*, δ-proteobacteria(leu), Dictyoglomi, *Thermomicrobium roseum*
methionine	Lactobacillales (exept streptococcus spp.), *Desulfotomaculum reducens*
proline	Bacillales, Desulfitobacterium hafniense, *Desulfotomaculum reducens*
cysteine	Bacillales, Enterococcus faecalis, *Clostridium acetobutylicum*, Dictyoglomi
histidine	some Lactobacillales
arginine	*Clostridium difficile*
threonine	Bacillus cereus, *Clostridium difficile*
serine	some FIRMICUTES
alanine	–
ASP, ASN	some FIRMICUTES
glutamine	*Clostridium perfringes*
glycine	–
lysine	–

In addition to two known tryptophan transporters, *yhaG* and *ycbK*, two new tryptophan transport systems were identified: *trpXYZ* (Peptococcaceae, *Streptococcus spp*., *Paenibacillus larvae*) and *yocR*(*yhdH*)(*Bacillus cereus*).

New large family of amino acid ABC transporters was characterized. In addition to previously described methionine ABC transporter *yusCBA* (Zhang *et al*., 2003) we found five new amino acid ABC transporters from this ABC transporter superfamily: *yqiXYZ*(*ARG*), *hisXYZ*(*HIS*), *yckKJI*(*CYS/MET*), *aspQHMP*(*ASP*), *ytmKLM*(*MET*).

The specificity of various possible amino acid permeases was predicted: *yvbW*(*LEU*), *ykbA*(*THR*), *lysX*(LYS), *RDF02391*(*ARG*).

Genes encoding transporters from branched-chain amino acid transporter family was found to be regulated by three amino acids: ILE (some Bacillales, Lactobacillales and Clostridiales), VAL(some Lactobacillales), THR (*Bacillus cereus, Clostridium tetani*).

Analysis of the methionine-specific T-box regulatory signals allowed us to identify hypothetical oligopeptide ABC transport system in Gram-positive bacteria, *opp*, which is possibly involved in the uptake of some methionine precursors or oligopeptides.

Table 2. Regulation of amino acid transporters by T-box antitermiantion in Gram-positive bacteria

Gene	Sp.	Predicted function	Bacteria
ycbK	TRP	tryptophan-specific permease	*Bacillus subtilis, Bacillus licheniformis*
yhaG	TRP	tryptophan-specific permease	Clostridiales
yvbW	LEU	leucine-specific permease	*Bacillus subtilis, Bacillus licheniformis*
ykbA	THR	threonine-specific permease	*Bacillus subtilis*
ybgF/aapA	?	?	*Lactobacillus reuteri*
yheL	TYR	Tyrosine transporter (Na+/H+ antiporter)	some Bacillales and Lactobacillales
lysX	LYS	lysine transporter	some Bacillales
	ILE	Branched-chain amino acid transporter family: ILE-specific	some Bacillales, Lactobacillales andClostridiales
brnQ_braB	THR	Branched-chain amino acid transporter family: THR-specific	*Bacillus cereus, Clostridium tetani*
	VAL	Branched-chain amino acid transporter family: VAL-specific	some Lactobacillales
yusCBA	MET	methionine ABC transporter	Lactobacillales, *Enterococcus faecalis*
yqiXYZ	ARG	arginine ABC transporter	*Clostridium difficile*
hisXYZ	HIS	histidine ABC transporter	Lactobacillales, *Clostridium difficile*, *Listeria monocytogenes*, *E. faecalis*
yckKJI	CYS	cysteine ABC transporter	*Clostridium acetobutylicum*
	MET	methionine ABC transporter	some Lactobacillales
aspQHMP	ASP	ASP(ASN) ABC transporter	*Lactobacillus johnsonii*
ytmKLM	MET	methionine ABC transporter	*Leuconostoc mesenteroides*
	TRP	TRP-*specific sodium dependent transporter*	*Bacillus cereus*
	PHE	PHE-*specific sodium dependent transporter*	*Bacillus cereus*
	LEU	LEU-*specific sodium dependent transporter*	*Bacillus cereus*
	?	*sodium dependent transporter*	*Clostridium tetani*
mtsABC opp	MET	uptake of unknown methionine precursors, possibly oligopeptides	some Lactobacillales
trpXYZ	TRP	tryptophan ABC transporter	Peptococcaceae, *Streptococcus spp.*, *Paenibacillus larvae*
RDF02391	ARG	arginine permease	*Clostridium difficile*
ABC-like	?	?	*Desulfotomaculum reducens*
CBX	?	?	*Clostridium botulinum*
gltT like	?	?	*some Clostridium spp.*

New possible amino acid transporters are in bold. Predicted specificity of an amino acid transporter is shown in second column.

AKNOWLEGEMENTS

This study was partially supported by grants CDF RBO-1268 from the Ludwig Institute for Cancer Research and 55000309 from the Howard Hughes Medical Institute.

REFERENCES

Grundy F.J., Rollins S.M., Henkin T.M. (1994) Interaction between the acceptor end of tRNA and the T box stimulates antitermination in the *Bacillus subtilis* tyrS gene: a new role for the discriminator base. *J. Bacteriol.*, **176**, 4518–4526.

Leontiev L.A., Lyubetsky V.A. (2006) Massive search of conserved regulatory structures containing T-boxes: results of calculation. *Information processes*, **6**, 20–23.

Panina E.M., Vitreschak A.G., Mironov A.A., Gelfand M.S. (2003) Regulation of biosynthesis and transport of aromatic amino acids in low-GC Gram-positive bacteria. *FEMS Microbiol. Lett.*, **28**, 211–220.

Rodionov D.A., Vitreschak A.G., Mironov A.A., Gelfand M.S. (2004) Comparative genomics of the methionine metabolism in Gram-positive bacteria: a variety of regulatory systems. *Nucl. Acids Res.*, **32**, 3340–3353.

Vitreshchak A.A., Mironov A.A., Gelfand M.S. (2001) Computer prediction of RNA secondary structure. The RNApattern program: searching for RNA secondary structure by the pattern rule. In *Proceedings of the Third International Conference "ComplexSystems: Control and modeling problems*". Russia, Samara, pp. 623–255.

Zhang Z., Feige J.N., Chang A.B., Anderson I.J., Brodianski V.M., Vitreschak A.G., Gelfand M.S., Saier M.H. Jr. (2003) A transporter of *Escherichia coli* specific for L- and D-methionine is the prototype for a new family within the ABC superfamily. *Archives of Microbiology*, **180**, 88–100.

RSCU_COMPARER: A NEW STATISTICAL TOOL FOR PRACTICAL ANALYSIS OF CODON USAGE

Vladimirov N.V.**, Kochetov A.V., Grigorovich D.A., Matushkin Yu.G.
Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia
* Corresponding author: e-mail: nikita@bionet.nsc.ru

Key words: RSCU, synonymous codons, t-test, gene expression, transgene

SUMMARY

Motivation: Many existing methods for statistical analysis of codon usage are efficient for deep investigation of codon usage patterns, but not for practical use in bioengineering. Practitioners are interested in simple and easy-to-use methods for determining the most unfavorable codons in a heterologous gene for its expression in a particular organism.

Results: A new statistical method for practical codon usage analysis is proposed. It allows a user to reveal statistically significant differences of codon usage patterns between two samples of highly and lowly expressed genes (provided by the user), and to determine the set of optimal and suboptimal codons that should be preserved or avoided in heterologous gene engineering. The novelty of the method is based on observation that distribution of $\log_{10}$(RSCU) values is close to normal (RSCU, relative synonymous codon usage). This transformation allows parametric statistics, such as Student's t-test, to be applied to comparison of codon usage patterns.

The program *RSCU_comparer* may be applied in practical bioengineering studies for selection of optimal and non-optimal codons in particular species using training sets of highly and lowly expressed genes.

Availability: http://wwwmgs.bionet.nsc.ru/mgs/systems/transgene/rscu.html.

INTRODUCTION

It is well known that the usage of synonymous codons in eukaryotic genes varies greatly in a species-specific manner. In many prokaryotic and in some eukaryotic organisms (e.g. *Saccharomyces cerevisiae*, *Drosophila melanogaster*) highly expressed genes are characterized by a preferential usage of more frequent synonymous codons. Currently, the relationship between translation elongation efficiency and preferences in the usage of synonymous codons in genes of most eukaryotic organisms is known poorly. However, an adjustment of the synonymous codon content is frequently used in gene engineering experiments to enhance a transgene expression rate. An increased content of species-specific synonymous codons preferentially used by highly expressed genes of a host organism may improve transgene expression characteristics.

METHODS AND ALGORITHMS

Training sets of highly (H) and lowly (L) expressed genes are provided by the user. For both samples the RSCU indices of codon usage bias (relative synonymous codon usage) (Sharp, Li, 1987) are calculated for 59 codons that have synonymous alternatives:

$$RSCU_i = \frac{X_i}{\frac{1}{n_i}\sum_{j=1}^{n_i} X_j}.$$

Here X_i is the number of *i*-th codon occurences in the (H) or (L) gene set, and n_i – the number of alternative synonymous codons for *i*-th codon (from 2 to 6).

In such a way, codon usage of (H) and (L) samples is represented by two 59-dimensional vectors: $RSCU^H = (x_1, \ldots, x_{59})$ and $RSCU^L = (y_1, .., y_{59})$, respectively.

The idea of the method is based on observation that distributions of $\log_{10}(RSCU)$ values are usually close to the normal distribution (see Examples). The log-normality of $RSCU^H$ and $RSCU^L$ allows to apply parametric statistics for comparison of codon usage patterns between (H) and (L) sets. For this purpose we used paired Student's t-test implemented in R statistical package (www.r-project.org). Since paired t-test requires normality of paired differences, at first we perform the Kolmogorov-Smirnov (K-S) test for normality of paired differences $D = \log_{10}(RSCU^H) - \log_{10}(RSCU^L) = (d_1, .., d_{59})$. For this purpose we use K-S test also implemented in R.

The vector $(d_1, .., d_{59})$ reflects the difference in codon usage frequencies between (H) and (L) samples. The smallest d_i values correspond to the suboptimal codons. Selection of suboptimal codons is performed by taking a lower Q_{sub}-quantile of d_i distribution. Codons whose d_i fall into the quantile limits are considered suboptimal. Avoidance of these codons in a heterologous gene may increase its translation elongation rate in the selected organism. In the same manner, the optimal codons (with high d_i values) are selected using the upper Q_{opt}-quantile (10 % by default).

EXAMPLES

The closeness of log10(RSCU) to normal distribution may be seen on the following example (Fig. 1). We selected three samples of genes from *E. coli* representing highly (High), lowly (Low) expressed and randomly selected (Rand) genes. Each set contains 20 genes. The High set consists of ribosomal genes, the Low set includes genes with the lowest elongation efficiency index (EEI) (Likhoshvai, Matushkin, 2002) that estimates codon bias, and Rand includes genes randomly selected using the RSA-tools (http://rsat.ulb.ac.be/rsat/).

The closeness of $\log_{10}(RSCU)$ distributions to normal is confirmed by the Kolmogorov-Smirnov test for normality:

High: $D_{KS} = 0.148$, $p = 0.164$; Low: $D_{KS} = 0.125$, $p = 0.309$; Rand: $D_{KS} = 0.0979$, $p = 0.623$.

The K-S test for normality of paired differences $(d_1,..d_{59})$ between High, Low and Rand samples:

(H-L) $D_{KS} = 0.150$, $p = 0.200$; (H-R) $D_{KS} = 0.164$, $p = 0.128$; (L-R) $D_{KS} = 0.0907$, $p = 0.716$.

After validation of normality of $(d_1, .., d_{59})$, the user can apply t-test for comparison of codon usage patterns:

1) High *vs* Low**:** $T = -2.68$, $p = 0.009$, samples are significantly different (threshold is 0.05).

2) High *vs* Random: $T = -3.58$, $p = 0.0007$, samples are significantly different.

3) Random *vs* Low: $T = 1.31$, $p = 0.19$, difference is non-significant.

The paired *t* test allows to discriminate between gene sets characterized by a considerable difference in the usage of synonymous codons (High *versus* Rand, High *versus* Low for the *E. coli* example). To confirm the applicability of our method to different organisms with known correlation between codon usage bias and gene expression, we analyzed samples "High" and "Rand" from *E. coli*, *B. subtilis*, and *H. enfluenzae* for positive examples, and from *H. pylori* for a negative one. In each organism

"High" gene set comprised 40 ribosomal genes, and "Rand" gene set consisted of 40 randomly selected genes. The results are shown in Table 1. The K-S test shows normality of all paired differences, and the t-test shows highly significant differences in codon usage of the bacteria except *H. pylori*, which is known to lack translational codon bias.

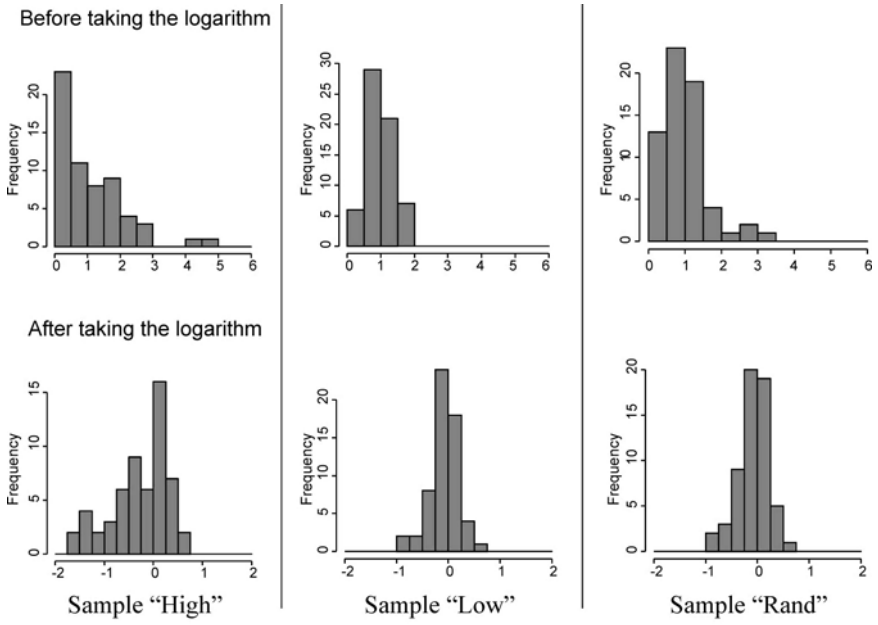


Figure 1. Distributions of RSCU and log10(RSCU) values (x-axis) for 3 sets from *E. coli*. The y-axis corresponds to the number of occurrences.

Table 1. Kolmogorov-Smirnov test for (d_1, .., d_{59}) normality, and the t-test for differences between "High" and "Rand" samples in bacteria

Organism	K-S test, p-values	t-test, p-values
E. coli	0.59	0.0006
B. subtilis	0.64	0.01
H. influenzae	0.56	0.009
H. pylori	0.31	0.9

IMPLEMENTATION

RSCU values are calculated with C-program, and transferred into R script that filters the data and computes statistics with t-test() and ks.test() functions.

The program takes 2 sets of coding sequences (CDS) assumed to be highly (H) and lowly (L) expressed in the organism of interest. CDS should be submitted in FASTA format.

The program computes:

1. K-S test for normality of (H), (L) and (H-L) $\log_{10}$(RSCU) distributions,
2. paired t-test for mean of RSCU differences between (H) and (L) samples,
3. codons completely absent in (H) set (if any),
4. the most unfavorable codons (defined by Q_{sub}),
5. the most favorable codons (defined by Q_{opt}).

User-defined parameters

1. p_{max}, a threshold of significance for *p*-values (default 0.05);
2. Q_{sub}, a quantile threshold for suboptimal codons selection (default 10 %);
3. Q_{opt}, a quantile threshold for optimal codons selection (default 10 %).

Limitations

In general, avoidance of the most unfavorable codons may be useful even if the difference between the (H) and (L) samples is not statistically significant, because the level of significance depends on the sample sizes and the selection procedure. The selection of appropriate "High" and "Low" expression gene sets may be an uneasy task because of a difference in tissue-specific expression rate (for multicellular organisms), possible difference between expression of members of multigene families, etc.

It should also be noted that in some organisms codon content may not correlate with translation level significantly (possibly because the rate-limiting stage is translation initiation, or for some other reasons).

DISCUSSION

The novelty of the method is based on observation that distribution of $\log_{10}$(RSCU) values is close to normal. This transformation allows application of parametric statistics, such as Student's t-test, to comparison of codon usage patterns. A similar approach is widely used in statistical analysis of microarray data (Stekel, 2003), where experimental data on gene expression are normalized (including log-transformation) and compared with reference samples. The (H) set may include genes for ribosomal proteins, heat-shock proteins, translation elongation factors, actins, tubulins or other abundant proteins. The (L) set may include genes encoding transcriptional factors, growth factors, receptors, protein kinases (Kochetov *et al.*, 1998) or randomly selected genes. Also, the (H) and (L) gene sets may be composed on the base of a microarray, SAGE or proteomic data. We recommend to use sets containing at least 20 genes for representative statistics.

RSCU_comparer may be applied in practical bioengineering studies for selection of optimal and non-optimal codons in particular species using training sets of highly and lowly expressed genes. It provides an opportunity to select unfavorable synonymous codons to be replaced for enhancing expression of a transgene CDS.

ACKNOWLEDGEMENTS

This work was supported by the Russian Foundation for Basic Research (grants Nos 05-04-48207, 06-04-49556), the Program of Russian Academy of Sciences (Dynamics of Plant, Animal, and Human Gene Pools), and the Program of RAS Presidium "Origin and evolution of biosphere" (Evolution of molecular-genetic systems: computer analysis and modeling) No. 10104-34/P-18/155-270/1105-06-001/28/2006-1. We thank SB RAS Complex Integration Program (No. 5.3) and Ministry of Industry, Science and Technologies of the Russian Federation (grant No. 2275.2003.4) for partial support.

REFERENCES

Kochetov A.V. *et al.* (1998) Eukaryotic mRNAs encoding abundant and scarce proteins are statistically dissimilar in many structural features. *FEBS Lett.*, **440**(3), 351–355.

Likhoshvai V.A., Matushkin Yu.G. (2002) Differentiation of single-cell organisms according to elongation stages crucial for gene expression efficacy. *FEBS Lett.*, **516**, 87–92.

Sharp P.M., Li W.H. (1987) The codon adaptation index – a measure of directional synonymous codon usage bias, and its potential applications. *Nucl. Acid Res.*, **15**, 1281–1295.

Stekel D. (2003) Microarray Bioinformatics. Cambridge University Press, Cambridge, 115–123.

ASSOCIATION STUDY OF SNP OF THE TNF-ALPHA GENE WITH BOVINE LEUKOSIS AND EVALUATION OF ITS FUNCTIONAL SIGNIFICANCE

***Yudin N.S.*[1*], *Vasil'eva L.A.*[1], *Kobzev V.F.*[1], *Kuznetsova T.N.*[1], *Ignatieva E.V.*[1], *Oshchepkov D.Yu.*[1], *Voevoda M.I.*[1,2], *Romaschenko A.G.*[1]**

[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia; [2] Institute of Internal Medicine, SB RAMS, Novosibirsk, Russia

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

Key words: SNP, TNF-alpha gene, bovine leucosis, association study, cattle, binding site recognition, transcriptional factor

SUMMARY

Motivation: The bovine leukemia virus (BLV) is an infectious agent that has taken enormous economic tolls in cattle breeding worldwide. Efficient search of the molecular-genetic markers of susceptibility to this infectious disease is required for effective breeding. The TNF-alpha protein plays an important defensive role during the early stages of the disease. The aim was to analyze the association of polymorphism of the TNF-alpha gene with bovine leucosis susceptibility and to evaluate its possible functional contribution to the pathogenesis of the disease.

Results: A single nucleotide polymorphism (SNP), G to A substitution in the first intron of the TNF-alpha gene (G4389A), was detected in black-spotted cattle. The occurrence frequency of the variant A of this polymorphism was significantly higher among the BLV carriers and animals at the terminal stage of leucosis than in the healthy controls.

The region containing the SNP G4389A comprised the potential sites for the transcriptional factor binding (NF-kB, E2F/DP1, SAP1). The DNA region was highly homologous and consequently conserved in bovine, human mouse and rat. Hence, a cause of the TNF-alpha gene association with bovine leucosis may be alteration of DNA interaction with transcription factor due to the nucleotide substitution at the G4389A.

INTRODUCTION

The bovine leukemia virus (BLV) is an infectious agent causing great economic losses. The data in the literature indicate that the certain cattle breeds are more susceptible to the BLV infection than others. However, to combat leucosis and to provide effective breeding for resistance to BLV infection, search of the molecular-genetic markers of susceptibility to the bovine leucosis is needed. The TNF-alpha gene can be involved in leucosis susceptibility. It is known that TNF-alpha performs an important defensive function in the early phase of BLV infection. When exogenous TNA-alpha is added to BLV infected cells *in vitro*, the expression of viral antigens is strongly suppressed (Meirom *et al*., 1997). TNF(-/-) mice are rendered more susceptible to an infection with BLV (Muller *et al*., 2003). The mean mRNA expression level for TNF-alpha is considerably higher in the spontaneously proliferating peripheral blood mononuclear cells (PBMCs) derived from BLV-infected cattle than in non-spontaneously proliferating PBMCs from normal cattle (Konnai *et al*., 2006).

Detection of the association between the gene polymorphism under study and a disease does not, as yet, mean that the gene polymorphism is the direct cause of disease. Other neighbor polymorphisms may perhaps be causative. To define the putative involvement of polymorphism in the emergence of disease, an evaluation of its functional significance is another requirement. The current bioinformatics tools allow predicting the potential functional significance of any DNA region.

Here our aim is to analyze the association between the polymorphism (G4389A) of the first intron of the TNF-alpha gene with bovine leucosis and to evaluate the possible contribution of the polymorphism to the pathogenesis of the disease.

METHODS AND ALGORITHMS

Based on serological and hematological assays, 5 groups of black-spotted cattle were differentiated: with terminal leucosis (N = 11), persistent lymphocytosis (N = 14), and BLV-carriers (N = 20) from the cattle farm "Novospasskoe", also BLV-carriers (N = 63) and healthy (BLV-free) controls (N = 28) from the cattle farm "Tulinskoe". The groups of BLV-carriers from the two farms did not significantly differ by genotype and allele frequencies. For this reason, the two groups were pooled as one BLV-carrier group (N = 83). Allele-specific PCR was utilized for DNA genotyping. To recognize the potential binding sites for the transcriptional factors (TF), we applied the SITECON method, based on analysis of the physicochemical and conformational DNA properties (Oshchepkov *et al.*, 2004).

RESULTS

In black-spotted cattle widely bred in Russia, single nucleotide polymorphism in the first intron of the TNF-alpha gene was detected. The polymorphism resulted from G to A substitution at position 4389 (GenBank accession number Z14137).

The results of analysis of the association between the SNP G4389A of the TNF-alpha gene with BLV-carriers and various stages of leucosis are summarized in Table 1. Variant A at position 4389 occurred significantly more frequently in BLV-carriers and in those with leucosis at its terminal stage, as compared with the healthy controls. The variant A frequency in cattle with persistent lymphocytosis was also higher than in the controls; however, the differences did not reach statistical significance.

Table 1. Genotype and allele frequencies of the SNP G4389A of the TNF-alpha gene in BLV-carriers, and at the two leucosis stages

	Genotype frequencies, n (%)			Allele frequencies, n (%)	
	G/G	G/A	A/A	G	A
BLV-carriers (N=83)[1]	58 (69.9 %)	24 (28.9 %)	1 (1.2 %)	140 (84.3 %)	26 (15.7 %)
Persistent lymphocytosis (N = 14)[2]	12 (85.7 %)	2 (14.3 %)	0 (0.0 %)	26 (92.9 %)	2 (7.1 %)
Terminal leucosis (N = 11)[3]	7 (63.6 %)	4 (36.4 %)	0 (0.0 %)	18 (81.8 %)	4 (18.2 %)
Healthy controls (N = 28)	27 (96.5 %)	1 (3.5 %)	0 (0.0 %)	55 (98.3 %)	1 (1.7 %)

[1] Differences from the controls are significant for the genotypes ($\chi^2 = 8.2$, $p < 0.02$) and for the alleles ($\chi^2 = 6.3$, $p < 0.02$); [2] Differences from the controls are insignificant; [3] Differences from the controls are significant for the genotypes ($\chi^2 = 4.9$, $p < 0.03$) and for the alleles ($\chi^2 = 4.6$, $p < 0.03$).

Using the SITECON method for the recognition of the TF binding sites, we analyzed a DNA region of 100 bp of the first intron of the TNF-alpha gene containing the SNP G4389A. Recognition of the binding sites was performed for TF of about 200 types. It was found that the 100 bp region contained a considerable number of potential TF binding sites. The G to A substitution had the strongest effect on the context-dependent

conformational and physicochemical DNA properties significant for the interaction with the NF-kB, E2F/DP1, SAP1 factors (Fig. 1). Multiple alignment of the bovine, human, mouse, and rat DNA fragments of the TNF-alpha gene demonstrated that the motif containing G4389A is highly conserved in these four species (Fig. 2).

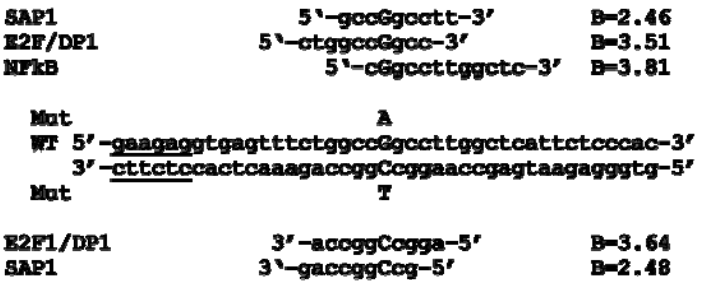


Figure 1. Results of context analysis of the intronic DNA fragment containing of SNP G4389A of the TNF-alpha gene. A rectangle outlines the bovine TNF-alpha gene sequence. The sequence that corresponds to the end of the first exon is underlined. The alternative nucleotides that occur at position 4389 are denoted by capital letters. The sequence regions of the bovine TNF-alpha gene that is homologous to the potential TF binding sites NF-kB, E2F/DP1, and SAP1 in direct orientation are above the gene sequence, those found in the reverse orientation are under it. B is the ratio of the scores for the two context similarities of the potential TF binding sites for the normal and rare alleles.

Figure 2. Multiple alignments of the bovine (GenBank accession number Z14137), human (AY066019), mouse (M20155) and rat (D00475) DNA fragments of the first intron of the TNF-alpha gene. The first exon is underlined. The identical nucleotides are shaded. SNP G4389A is in bold.

DISCUSSION

The obtained data indicate that the variant A of SNP G4389A affects the cattle susceptibility to infection with the BLV and also the clinical pattern of leucosis. There was no significant association between the polymorphism under study and persistent lymphocytosis presumably because of the small sample size.

The data on context DNA analysis for the first intron of the TNF-alpha gene and the high homology of the bovine nucleotide sequence comprising polymorphism with the same sequences in the three mammalian species demonstrate that the SNP G4389A occurs most likely in the functionally significant TF binding site or in a composite element. It is known that the NF-kB TF can regulate the expression of the human MICA (Molinero *et al*., 2004), and BCL3 (Ge *et al*., 2003) genes by binding to the specific sequence in their introns. Possibly, change in the ability of NF-kB to bind to the first intron of the A variant carriers may affect on the expression of the TNF-alpha gene and, as a consequence, modify antiviral immunity. Other alternatives may be the influence on the binding of the heteromeric E2F1/DP1 TF that regulates cell cycle or on the binding of the SAP1 TF that alters the expression of the *c-fos* oncogene. It should be stipulated that we detected an association between the SNP G4389A and BLV susceptibility at the early stage of the disease (BLV-carriers). This lends credibility to our suggestion that precisely NF-kB binding may be the cause of the association.

It should be noted that in human, mouse and rat, adenine is at position homologous to SNP G4389A, i.e. the rare variant that is associated with cattle susceptibility to infection with BLV and bovine leucosis. Search in the dbSNP database also detected no SNPs in a region 10 bp upstream and 10 bp downstream from SNP G4389A in human, mouse and rat. This was expected because in the closely related species as human and orangutan scanning on the long arm of the X chromosome also did not reveal shared SNPs between the species (Miller, Kwok, 2001). BLV is a model for studying of human leukemias caused by the closely related human T-cell leukemia virus type 1 (HTLV-1) (Johnson *et al*., 2001). There are reports indicating that BLV has potential infectivity and oncogenicity for humans (Johnson, Griswold, 1996, among other). Our current data suggests that the TNF-alpha gene is promising for the identification of hereditary predisposition to leukemia in human.

It may be concluded that the variant A of the TNF-alpha gene is associated with cattle susceptibility to infection with BLV and bovine leucosis. The current study has demonstrated this for black-spotted cattle. A mechanism underlying this association may possibly be the effect of polymorphism on the interaction between DNA and the NF-kB, E2F/DP1, and SAP1 transcriptional factors. This ultimately results in altered expression of the TNF-alpha gene.

ACKNOWLEDGEMENTS

Work was supported by the program "Dynamics of the gene resources of plant, animals and human" of the Russian Academy of Science and by the innovation project of Federal Agency of Science and Innovation IT-CP.5/001 "Development of software for computer modeling and design in postgenomic system biology (system biology *in silico*)". The authors are grateful to N.S. Ufimtseva for assistance in collection of bovine blood samples. They are also grateful to V.A. Beliavskaia and P.N. Smirnov for providing them the results of serological and hematological assays.

REFERENCES

Ge B. *et al.* (2003) NF-kappa B regulates BCL3 transcription in T lymphocytes through an intronic enhancer. *J. Immunol*., **171(**8**)**, 4210–4218.

Johnson E.S., Griswold C.M. (1996) Oncogenic retroviruses of cattle, chickens and turkeys: potential infectivity and oncogenicity for humans. *Med. Hypotheses*, **46(**4), 354–356.

Johnson J.M. *et al.* (2001) Molecular biology and pathogenesis of the human T-cell leukaemia/lymphotropic virus type-1 (HTLV-1). *Int. J. Exp. Pathol*., **82(**3), 135–147.

Konnai S. *et al.* (2006) Tumor necrosis factor-alpha up-regulation in spontaneously proliferating cells derived from bovine leukemia virus-infected cattle. *Arch. Virol*., **151(**2), 347–360.

Meirom R. *et al.* (1997) Levels and role of cytokines in bovine leukemia virus (BLV) infection. *Leukemia*, **11(**3), 219–220.

Miller R.D., Kwok P.Y. (2001) The birth and death of human single-nucleotide polymorphisms: new experimental evidence and implications for human history and medicine. *Hum. Mol. Genet*., **10(**20), 2195–2198.

Molinero L.L. *et al.* (2004) NF-kappa B regulates expression of the MHC class I-related chain A gene in activated T lymphocytes. *J. Immunol*., **173(**9), 5583–5590.

Muller C. *et al.* (2003) Lack of TNF alpha supports persistence of a plasmid encoding the bovine leukaemia virus in TNF(-/-) mice. *Vet. Immunol. Immunopathol*., **92(**1/2), 15–22.

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

PHYLOGENETIC CHANGES IN CHLOROPLAST GENOMES

Zotov V.S.[1], Punina N.V.[1], Dorokhov D.B.[1], Schaad N.W.[2], Ignatov A.N.[*1]
[1] Centre "Bioengineering", RAS, Moscow, Russia; [2] FDWSRU, USDA-ARS, MD, USA
[*] Corresponding author: e-mail: ignatov@biengi.ac.ru

Key words: evolution, chloroplast, adaptation, phenotype, computer analysis

SUMMARY

Motivation: Evolution of green plants cross a few major adaptation boundaries: land plants vs. algae, seed plants (*Embryophyta*) vs. others, and monocotyledons (*Liliopsida*) vs. dicots (*eudicotyledons*). Evolution of chloroplasts – symbiotic bacteria inside eukaryote cells, is different from evolution of nuclear genome in both rate and driving forces of phylogenetic modifications. Several distinct approaches were applied to evaluate an impact of evolutionary leaps on chloroplast genome content and composition in aim to connect structural changes to adaptive advantages of progressive phylogenetic lineages in green plants.

Results: A comparison of nucleotide and amino acid similarity for conservative genes, gene content, intron presence, amino acid and codon usage, frequency and strand asymmetry of oligonucleotides, and average distance between mononucleotides was made for 51 complete chloroplast genomes of 53 published to the date. The hypothesis of "quantum leaps" in chloroplast genome organization was tested by comparison of different parameters with consensus evolutionary distances obtained by nucleotide and amino acid similarity for conservative ribosomal rRNA and protein genes. The result demonstrates that transaction from algae to land plants, origin of seed plants and monocotyledons was attributed by significant changes in codon usage, frequency of asymmetric oligonucleotides and average similar nucleotide distance.

Availability: http://www.bionet.nsc.ru/bgrs2006/.

INTRODUCTION

Green plant chloroplast is a symbiotic prokaryote acquired by ancestor of green plants (Buetow, 1976), closely related to Cyanobacteria (fam. Nostocaceae). Evolution of chloroplasts and other parasitic and symbiotic bacteria have many striking similarities (Eremeeva *et al*., 2005), including massive loss of genes and reduction of GC content. Generally, changes of chloroplast genome correlate to evolution of nuclear one, with some exception (Chu *et al*., 2004). Unfortunately, methods of comparison applied for chloroplast genes rarely considered conventional bioinformatics approaches for analysis of genome composition developed for bacterial genome (Wolf *et al*., 2001). Driving forces of bacterial evolution are different from ones in complex eukaryote genome. Gene gain/loss and changes in DNA composition in bacteria substitute more complex regulatory mechanisms in the "host" nuclear genome. Evolution of green plants cross a few major adaptation boundaries: land plants vs. algae, seed plants (*Embryophyta*) vs. others, and monocotyledons (*Liliopsida*) vs. dicotyledons (*eudicotyledons*), and those macro-evolutionary events can be traced in organization of such symbiotic organism as chloroplast.

METHODS

We compared 51 published complete chloroplast genome using nucleotide and amino acid similarity for gene content, amino acid and codon usage, frequency and strand asymmetry of oligonucleotides, and average distance between similar nucleotides (Zotov, unpublished). Genetic distances obtained for different criteria ways were compared to consensus matrixes for conservative ribosomal *16S-23S-5S rRNA* and ribosomal protein-coding genes using Mantel's test (Manly, 1985). Genetic distance matrixes for the used genes were calculated using Tamura-Nei Model for nucleotides and Poisson Model for amino acid sequences (Nei, Kumar, 2000).

Significant differences between pooled phylogenetic close by the ribosomal genes genomes were considered as evidence of novel mechanism of adaptation of chloroplast genome to beneficial properties of "host" organism. Tree major evolutionary events were tested: conquest of land by algae, origin of seed plants, and separation of monocotyledons. A phylogenetic tree was built by Ward's algorithm for Pearson's distance (Fig. 1) to illustrate the significant leaps in asymmetry of frequent octamer repeats, and principal component analysis was used to illustrate codon usage differences (Fig. 2). Both analyses were made in STATISTICA 6.0® (StatSoft, USA).

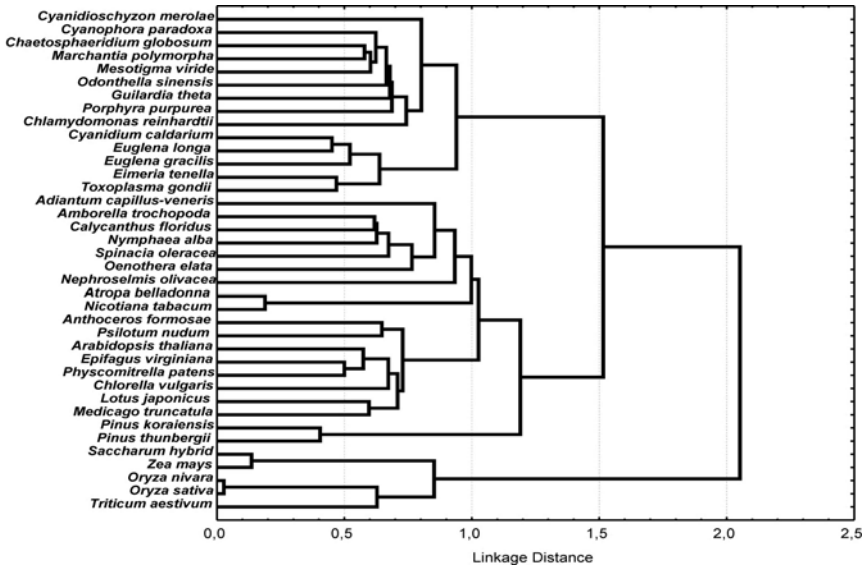


Figure 1. Ward's tree based on Pearson's distance between asymmetry of frequent octamer repeats in 40 complete chloroplast genomes. Arrow indicate a major junction between algaes, dicotyledons and monocotyledons.

RESULTS

Summarized results of comparison are showed in Table 1. Evolutionary independent gene loss was observed at different branches of the phylogenetic tree of chloroplasts. The major evolutionary event in plant life – transition from algaes to land plants was attributed by significant changes in strand asymmetry of frequent oligonucleotides (Fig. 1), when appearance of seed plants – by codon usage (Fig. 2). The most significant changes were found in chloroplast of monocotyledons – amino acid and codon usage, strand asymmetry of oligonucleotides and average distance between similar mononucleotides were remarkably distinct in this highly progressive but genetically narrow group.

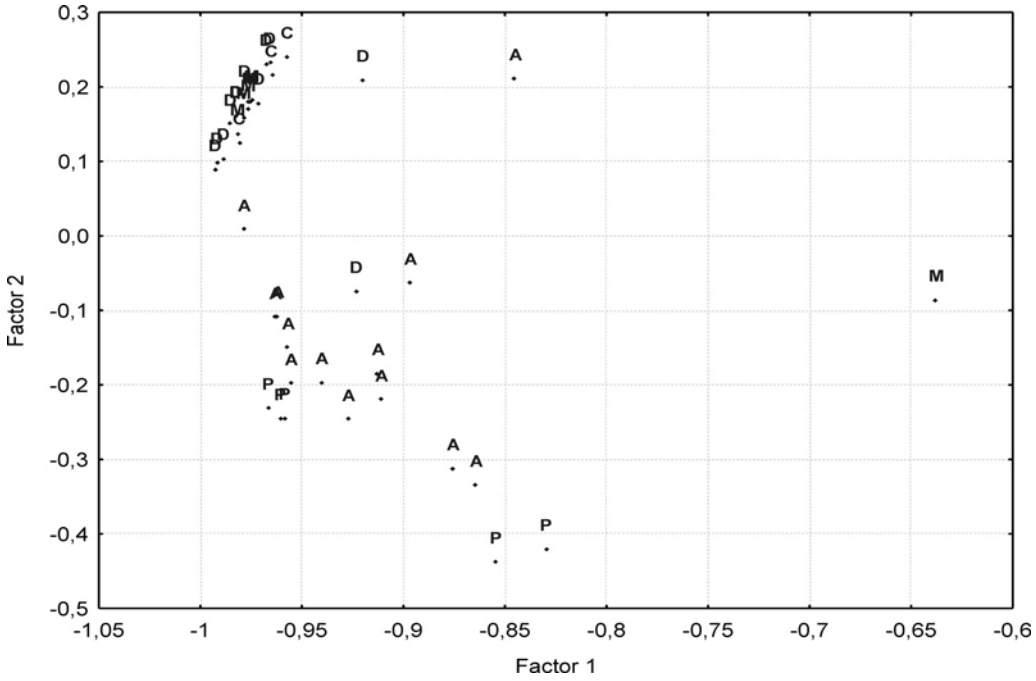


Figure 2. Principal Component Analysis of codon usage in 51 chloroplast genome. A – algaes, P – parasitic plants, C – conifers; D – dicotyledons; M – monocotyledons. The first and the second factors explained 54 % of total variation.

Table 1. Significant differences in matrixes of genetic distances between pooled chloroplast genomes

Parameter	Pooled comparison*	Difference	p-value
Gene content	1	N	na**
	2	N	–
	3	N	–
Amino acid usage	1	N	na
	2	N	–
	3	Y	< 0.01
Codon usage	1	N	na
	2	Y	< 0.01
	3	Y	< 0.001
Frequency of oligonucleotides	1	N	na
	2	N	–
	3	N	–
Strand asymmetry of oligonucleotides	1	Y	< 0.01
	2	N	na
	3	Y	< 0.001
Average distance between similar mononucleotides	1	N	na
	2	N	–
	3	Y	< 0.001

* Pooled comparison: 1, Algae vs. land plants; 2, Seed plants and conifers vs. others; 3, monocotyledons vs. others; ** na, not applicable.

CONCLUSIONS

The hypothesis of "quantum leaps" in chloroplast genome organization was supported by several parameters of the complete genome sequences. The obtained results demonstrate that evolution of chloroplasts led to significant and rapid changes in usage of codons and

frequency of oligonucleotides that can be indicative for more strict control of chloroplast gene expression from nuclear genes for better adaptive reaction of the whole plant.

ACKNOWLEDGEMENTS

The work was supported in part by the grant 3431 of the International Science and Technology Center.

REFERENCES

Buetow D.E. (1976) Phylogenetic origin of the chloroplast. *J. Protozool*., **23**(1), 41–47.

Chu K.H., Qi J., Yu Z.G., Anh V. (2004) Origin and phylogeny of chloroplasts revealed by a simple correlation analysis of complete genomes. *Mol. Biol. Evol*., **21**(1), 200–206.

Eremeeva M.E., Madan A., Shaw C.D., Tang K., Dasch G.A. (2005) New perspectives on rickettsial evolution from new genome sequences of rickettsia, particularly *R. canadensis*, and *Orientia tsutsugamushi*. *Ann. N Y Acad Sci*., **1063**, 47–63.

Manly B.F.J. (1985) *The Statistics of Natural Selection*. Chapman and Hall, London, 484 pp.

Nei M., Kumar S. (2000) Molecular Evolution and Phylogenetics. Oxford University Press, New York.

Wolf Y.I., Rogozin I.B., Grishin N.V., Tatusov R.L., Koonin E.V. (2001) Genome trees constructed using five different approaches suggest new major bacterial clades. *BMC Evol. Biol*., **1**, 8.

PART 6. OTHER TOPICS RELATED WITH BIOINFORMATICS

INVESTIGATION OF NAMED ENTITY RECOGNITION IN MOLECULAR BIOLOGY BY DATA FUSION

***Arrigo P.*[*1], *Cardo P.P.*[2]**
[1] CNR Institute for Macromolecular Studies, Genova, Italy; [2] Dept. of Health Science, University of Genoa, Genova, Italy
[*] Corresponding author: e-mail: arrigo@ge.ismac.cnr.it

Key words: text mining, semantic analysis, concept formation, molecular interaction

SUMMARY

Motivation: The amount of published scientific literature is fast expanding its management and processing is become a burden task. Text Mining (TM) is acquiring a key role for bioinformatics; it seems one of more suitable approaches for heterogeneous data sources integration. Textual data has been recently used to support scientific hypotheses generation (Literature Based Discovery). In this work we have considered the screening of previous unknown molecular in a literature based discovery perspective. The identification of molecular species, that could interact among them, is the first step for *in silico* design of molecular interaction networks. In order to achieve this goal, we have need to extract the more standardized set of potentially interacting molecules. The published articles reflects the fragmentation biomedical researches, this situation could affect the reliability of the set.

During the temporal evolution, new published papers can modify the knowledge about molecular interactions. The evaluation these changes on knowledge is important in a model development perspective. The screening of potentially interacting molecules could be considered equivalent to linguistic *named entity recognition* process. In this paper we have applied an ensemble of unsupervised learning machines to selection and extraction of *named entities* associated to potentially interacting molecules; the analysis has been focused on the changes emerged in PubMed repository during the period of time 1985–2000.

Results: A set of PubMed queries has been analyzed; everyone of which was a molecular entity. Each corresponding set of PubMed abstracts has been separately retrieved and processed; the retrieval phase has limited to the period 1985–2000. Each set has been split into three chunks; each chunk represent a five year sub interval. This procedure allowed us to screen *named entities*, specific for each time interval, associated with potentially interacting molecules; The recognition of time invariant named entities is essential for subsequent molecular interaction screening. A data-fusion system, based on self-organization paradigm, seems to be able to evaluate the temporal modification in textual information. Our system has detected, in this preliminary analysis, several named entities that can be functionally related with the original query.

Availability: http://biocomp.ge.ismac.cnr.it/

INTRODUCTION

The development of new systems for automated scientific hypothesis generation is one of more important challenge for Machine Learning. The automated scientific hypotheses generation could give an important contribution to biomedical researches. The biomedical researches shows a very high level of specialization and, consequently, the knowledge is

often fragmented. The difficulty to obtain of a general landscape, about a specific research topic, can limit the real knowledge advancement and the weakness of connection among different disciplines does not allow to easily correlate their results. The analysis of textual information could be a suitable approach to improve data integration among different disciplines. The exploitation of biomedical literature has urged the development of efficient informatics tools to organize and analyze this huge amount of unstructured data. The Natural Language Processing (NLP) methods has been recently used for several bioinformatics applications, such as protein fold recognition (MacCallum *et al*., 2000) or microarray analysis validation (Chagoyen *et al*., 2006). More recently, some authors have proposed the integration of NLP method into a Machine Learning system for automated scientific hypotheses generation (Weeber *et al*., 2005). The hypotheses generation, based on textual information, is quite different from the common rule generation methods such as inductive logic programming (ILP). The literature based discovery (LBD) is strongly dependent on the choice of semantic approach. Different semantic approaches are available: the connectionism is one of them. The Connectionist models (Artificial Neural Networks) are suitable for a large variety of applications, basically based on their statistical properties. In fact the origin of ANN was psychological (neuroinformatics); some neural models has been developed to simulate specific cognitive tasks (Hinton, 1990). In a knowledge discovery perspective, the more important property of connectionist systems seems to be their capability of bridging together symbolic and semantic level (Doursat, Petitot, 2005). The psychological roots of ANN could be useful to investigate the dynamical process of concept formation (hypotheses formulation).

METHOD

The basic module of the system has been described in a previous paper (Fattore, Arrigo, 2005). A data-fusion process combines multiple data sources and multiple classifier in order to produce valuable information for the user. In this application the multiple sources are the different time-specific set of PubMed Abstracts. The multiple classifier is constituted by an ensemble of self-organizing basic modules. Different approaches are available for information extraction and analysis: the combination of classifiers is one of them (Giacinto *et al*., 2000). We have chose a set based *voting* approach to merge the classifiers results. The analysis of temporal modification in textual data could be considered as a possible application for data-fusion process. The system is constituted by a set of unsupervised networks connected to the *voting* module (Anderson, 1999). These architecture seem to be the more suitable method to investigate the emergence of new *named entities* (NE) in the different time intervals. A *named entity* is a single word, or a sequence of words, that univocally identify a linguistic object; in molecular biology protein and gene names are *named entities*. We have identified these linguistics structures as Molecular Named Entities (MNE). A MNE is an acronym or composite word, that identify molecular species, as biological as chemical. The aim of this work is to investigate the changes in MNE recognition during temporal research advancement. The analysis has been performed on a 15 years time interval (1985–2000) splitting this period into three sub interval. We have indicated Q_{Δ_t} the query for the time interval Δ_t and $D(Q(\Delta_t))$ the corresponding set of retrieved abstracts. We have defined the set of trainable classifiers $C = \{C_1, \cdots, C_{\Delta_t}\}$. Each C_{Δ_t} has been singularly trained with a $D(Q(\Delta_t))$. When the learning phase terminated, we have extracted a *bag-of-word* (BOW) from each classification module. A *bag-of-word* is, roughly speaking, a set of words. The elements of this kind of list are not classified taking into account linguistic categories (noun, verb, adjective). In our case the BOW is the union set of t terms

extracted from representative vectors (t_{p_Δ}) of each classifier *k*. In order to take into account acronyms and composite word we have developed a specific routine to calculate the relative distance between two entities; it has been computed as spatial separation between the two entities in the phrase they co-occur. This distance permits to recognize composite MNE such as TGF-beta or beta-amyloid. The prototype of the system will be available at http://biocomp.ge.ismac.cnr.it. The new prototype will substitute the currently accessible PubClust system. The current available multiple classifier system has been developed by using C++ and, for the test it is running under Windows XP operating system. The current version of prototype is working only on biomedical abstract and is cannot, at the moment, screen chemical named entities.

RESULTS AND DISCUSSION

A biological heterogeneous set of query has been used, each one identified a specific MNE. These entities has been selected taking into account the PubMed download constrains. The system cannot retrieve from PubMed more than 5000 abstracts. Table I summarizes the MNE recognized by the system for each predefined time interval. We have obtained a reduced set of molecular named entities; several MNE listed in the table could be considered trivial solutions but they are not trivial for an artificial learning system. This work was essentially addressed to study of the effect of learning dynamic of a multiple self-organizing classifier during the MNE extraction phase.

Table 1.

Query	Molecule that potentially interact with the query		
	1985–1990	1990–1995	1995–2000
EGFR	EGF, tyrosine-kinase	tyrosine-kinase, EGF	tyrosine-kinase, EGF
Angiotensin I	Ang, captopril, renin	captopril, bradikinin, AT1, renin	renin, Ang, AT1, AT2, Losartan, bradikinin
VEGF	VEGF(*), dexamethasone, VPF	VPF, VEGF(*), KDR tyrosine-kinase	kinase
Homocysteine	homocysteine(*), methionine	folate, vitamin, cysteine, homocysteine(*)	folic-acid, B12, vitamin, reductase
PDGF	EGF, PDGF(*)	PDGF(*), EGF, tyrosine-kinase, kinase, TGF-beta	kinase, PDGF(*), EGF, tyrosine-kinase
Sp1	DNA, cDNA, RNA	cDNA, DNA, mRNA, TATA	DNA,TATA
TATA-box	mRNA, cDNA	mRNA, RNA, DNA, TBP, Sp1	TBP, RNA, DNA
c-myc	mRNA, RNA, DNA, myc	RNA, DNA	myc, DNA, p53, kinase, mRNA
Histone H4	DNA, kinase, mRNA, RNA, H2A, H2B	DNA, kinase, mRNA, H2A, H2B	DNA, mRNA

Our analysis pointed out some interesting results. For instance the MNE H2A and H2B seem to lose their importance regarding the query *histone H4*. Regarding Angiotensin I, the *captopril* drug reduced its importance, on the opposite *losartan* drug acquired importance during the time evolution. Other pharmacological active molecules has been recognized: *dexamethasone* and *Ang*, an active heptapeptide obtained from Angiotensin I. In the case of *c-myc*, the p53 and kinase MNE has emerged as potential interacting molecules. In addition the result summary evidences the functional relation between tyrosine-kinase and EGFR. This interaction is quite well established, instead the potential interaction between tyrosine-kinase and PDGF seems to be more recently

screened. A reliable MNE recognition is fundamental to identify the context in which they are embedded. The MNE are the basic tag to select the anchor phrases that could contain information about molecular interactions. The current version of the system does not allows to analyze the NE associated to chemical compounds, that can have great relevance in the investigation of protein drug interactions; in order to integrate the chemical compound screening we are developing a specific module for the recognition of entities of chemical compounds.

ACKNOWLEDGEMENTS

The work was supported by the EC STREP project CARDIOWORKBENCH LSHB-CT-2005-018671.

REFERENCES

Anderson B. (1999) Kohonen neural networks and language. *Brain Lang.*, **70**, 86–94.

Doursat R., Petitot J. (2005) Dynamical systems and cognitive linguistics: toward an active morphodynamical semantics. *Neural Netw.*, **18**, 628–638.

Fattore M., Arrigo P. (2005) Knowledge discovery and system biology in molecular medicine: an application on neurodegenerative diseases. *In Silico Biol.*, **5**, 199–208.

Giacinto G., Roli F., Fumera G. (2000) Design of effective multiple classifier system for cluster analysis. *Proceedings of 15th IEEE Conference on Pattern Recognition*, **2**, 160–163.

Hinton G. (1990) (ed.) *Connectionist Symbol processing.* Bradford Book, MIT Press, Cambridge, Massachusetts.

MacCallum R.M., Kelley L.A., Sternberg M.J. (2000) SAWTED: structure assignment with text description – enhanced detection of remote homologues with automated SWISS-PROT annotation comparisons. *Bioinformatics*, **16**, 125–129.

Weeber M., Kors J.A., Mons B. (2005) Online tools to support literature-based discovery in the life sciences. *Brief Bioinform.*, **6**, 277–286.

GRAPH THEORY ALGORITHM FOR SOLUTION OF COMPUTATIONAL PROBLEMS OF GENE MAPPING

Axenovich T.I.
Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia
e-mail: aks@bionet.nsc.ru

Key words: genetic mapping, pedigree, loop breaking, computer simulation

SUMMARY

Motivation: Pedigrees coming from isolated populations are very informative for gene mapping of complex traits due to their low genetic heterogeneity. Usually these pedigrees contain a lot of loops, which make impossible an application of likelihood-based methods of linkage analysis. The loops have to be broken to calculate likelihood using approximate methods. Optimal selection of breakers allows us to reduce the loss of genetic informativity. Existent algorithms of breakers selection are not effective for pedigrees from isolated populations where many individuals are unobserved and large number of loci are analysed.

Results: We used classical graph-theory Kruskal approach for selection of optimal loop breakers. This algorithm needs to define weights of edges for pedigree-graph. We propose to estimate edge weight by relative loss of mean relationship after the elimination of this edge from the graph. We estimated the loss of pedigree informativity in a series of inbred families with hundreds of loops under different algorithms of loop breaking and demonstrated that our algorithm provides minimal loss of information. In addition we compared power of linkage analysis of a pedigree with multiple loops using our and alternative methods for simplification of pedigree structure and demonstrated that our method is characterised by higher power.

Availability: We implemented our method in a software LOOP_EDGE. It is free available over the Internet on http://mga.bionet.nsc.ru/nlru/.

INTRODUCTION

Pedigrees coming from isolated populations present the unique material for gene mapping of complex traits because of their low genetic heterogeneity here. However a recurrent inbreeding in these populations results in a very complex structure of pedigrees coming from these populations. The pedigrees usually consist of many generations and contain a lot of loops. Many likelihood-based methods of linkage analysis can not be applied for the analysis of such pedigrees because the algorithms for likelihood calculation are sensitive to the presence of pedigree loops.

There are some approaches to calculation of likelihood for pedigrees with loops. Large part of them are based on cutting loops by “duplication” or “cloning” some of the participating individuals (the loop breakers). For pedigrees with small number of loops the exact likelihood calculation is possible (Ott, 1976). In case of multiple loops, only approximate methods of likelihood calculation are to be used. Usually they are based on conditioning of likelihood of zero-loop pedigree on phenotypes of copied individuals (Stricker *et al*., 1995; Wang *et al*., 1996).

Several variants of zero-loop pedigrees may be constructed from a large pedigree with multiple loops by selection of different sets of breakers. Some algorithms have been developed for selection an optimal set on the base on different objective variables, when for example, number of breakers (Becker *et al*., 1998) or number of breaker's possible genotypes (Vitezica *et al*., 2004) was minimized. The latter algorithm is especially popular (Lange, Goradia, 1987; O'Connell, Weeks, 1999). It is helpful when pedigree contains small number of loops and exact likelihood may be calculated. However if the pedigree contains a lot of unobserved members or many marker loci are under analysis, this algorithm becomes ineffective.

The aim of this study is the development of alternative algorithm for automatic selection of loop breakers in large pedigrees with great number of unobserved individuals suitable for multipoint linkage analysis.

ALGORITHM

We used the classical graph-theory Kruskal approach (Kruskal, 1956, 1997) for selection of optimal loop breakers. If the weights of all graph edges are known, the algorithm allows us to transform automatically the graph with multiple cycles into graph-tree with maximal weight of the edges included in the new graph. When this algorithm is applied to pedigree graph it guaranties the optimal selection of loop breakers. The problem is how to define the weights of all potential breakers.

It is known that informativity of a pedigree for genetic analysis depends on the relationship between its members. The mean relationship coefficient between pairs of genotyped relatives is one of the measures of pedigree informativity.

Let us denote the mean relationship coefficients in actual pedigree with multiple loops and in a pedigree constructed from the actual one by breaking a loop via coping of individual i as R_0 and R_i correspondingly. Then the weight of edge associated with potential breaker i may be defined as relative difference of these values $W_i = (R_0–R_i)/R_0$. The sum of edges corresponding to optimal set of breakers should be minimal.

PROPERTIES OF METHOD

We estimated the properties of our algorithm in two comparisons. At first we selected the loop breakers in several animal pedigrees with hundreds of loops using two measures of edge weight: mean relationship coefficient and number of offspring. We compared mean relationship coefficients in different zero-loop pedigrees constructed on the base of two measures of edge weight. For all analyzed pedigrees our algorithm gave the coefficient approximately 1.5 times higher than alternative method did.

Then we analyzed power of model-free parametric linkage analysis performed on different zero-loop pedigree structures: 1) single pedigree constructed by our method and 2) a set of the non-overlapping zero-loop fragments of initial pedigree which usually are constructed for existent computer programs. We performed the following experiment. The quantitative traits and the set of marker genotypes have been simulated on the base of initial pedigree with multiple loops. Then the simulated values have been ascribed to the corresponding individuals in zero-loop pedigrees and maximal *lod score* has been estimated for different zero-loop structures. The results demonstrated that zero-loop pedigree produced by our algorithm showed the values of *lod score* at average 0.25 higher then alternative zero-loop structure. Thus we can conclude that our algorithm has two advantages. It cuts the loops automatically and allows to increase the power of parametric linkage analysis. We implemented our method in a software LOOP_EDGE. It is free available over the Internet on http://mga.bionet.nsc.ru/nlru/.

ACKNOWLEDGEMENTS

This research was supported by the joint grant from the Netherlands Organization for Scientific Research (NWO) and the Russian Foundation for Basic Research (RFBR) and Research Program of RAS "Dynamics of Genomes".

REFERENCES

Becker A., Geiger D., Schaffer A.A. (1998) Automatic selection of loop breakers for genetic linkage analysis. *Hum. Hered*., **48**, 49–60.

Kruskal J. (1956) On the shortest spanning subtree of a graph and the traveling salesman problem. *Proc. Amer. Math. Soc*., **7**, 48–50.

Kruskal J. (1997) A reminiscence about shortest spanning trees. *Archivum Mathematicum (BRNO)*, **33**, 13–14.

Lange K., Goradia T.M. (1987) An algorithm for automatic genotype elimination. *Am. J. Hum. Genet*., **40**, 250–256.

O'Connell J.R., Weeks D.E. (1999) An optimal algorithm for automatic genotype elimination. *Am. J. Hum. Genet*., **65**, 1733–1740.

Ott J. (1976) A computer program for linkage analysis of general human pedigrees. *Am. J. Hum. Genet*., **28**, 528–529.

Stricker C., Fernando R.L., Elston R.C. (1995) An algorithm to approximate the likelihood for pedigree data with loops by cutting. *Theor. Appl. Genet*., **91**, 1054–1063.

Vitezica Z.G., Mongeau M., Manfredi E., Elsen J.M. (2004) Selecting loop breakers in general pedigrees. *Hum. Hered*., **57**, 1–9.

Wang T., Fernando R.L., Stricker C. (1996) An approximation to the likelihood for a pedigree with loops. *Theor. Appl. Genet*., **93**, 1299–1309.

BIOINFORMATICS EDUCATION IN THE INSTITUTE OF BIOMEDICAL CHEMISTRY RAMS: COURSE «BIOINFORMATICS – THE WAY FROM GENE TO DRUG» AND SPECIAL COURSE «BIOINFORMATICS AND COMPUTER-AIDED DRUG DESIGN»

Ivanov A.S.*[*], *Poroikov V.V.*, *Archakov A.I.
V.N. Orechovich Institute of Biomedical Chemistry, RAMS, Moscow, 119121, Russia
[*] Corresponding author: e-mail: alexei.ivanov@ibmc.msk.ru, ivanov@ibmh.msk.su

Key words: bioinformatics, education, computer-aided drug design, from gene to drug

SUMMARY

Motivation: Last years the genomics and proteomics discovery lead to extremely fast growth of information in molecular biology area. As bioinformatics becomes integral to biomedical research, there is a need of reorganization of education process in biomedical high schools by introducing bioinformatics theoretical and training courses.

Results: Described bioinformatics education is an example of integration of the higher school (Russian State Medical University, Medico-Biological Faculty) and academic science (Institute of Biomedical Chemistry RAMS). Education consists of theoretical course "Bioinformatics – the way from gene to drug" targeted for students on their 8th semester and special course «Bioinformatics and computer-aided drug design» for students on 11th semester. The last one includes theoretical study (lectures and seminars) and practical training in computer classroom as well as in laboratories of Bioinformatics Department of the Institute of Biomedical Chemistry RAMS.

Availability: All students from biomedical high schools as well as postgraduate students are welcome to our course "Bioinformatics – the way from gene to drug" which is giving since 2000 and special course "Bioinformatics and computer-aided drug design" which is giving since 1995. Lectures and seminars presentations (pdf format) are available on request from the authors.

INTRODUCTION

The fast growth of genomic and proteomic researches has led to multiple increase of our knowledge and explosive development of bioinformatics technologies, which become now an integral discipline in biomedical science. This situation generates the need of reorganization of education process in biomedical high schools by introducing bioinformatics theoretical and training courses.

Described bioinformatics courses are an example of integration of higher school (Russian State Medical University, Medico-Biological Faculty) and academic science (Institute of Biomedical Chemistry RAMS).

Students study algorithms and principles of applied programs for analysis of primary, secondary and tertiary structure of proteins; designing small ligands and simulation of their interaction with biological macromolecules; prediction of pharmacological properties of new substances; molecular modeling, molecular docking, database mining

and *de novo* design. Practical training is carried out in well-equipped computer classroom and in laboratories of the Institute of Biomedical Chemistry RAMS.

Bioinformatics education consists of theoretical course "Bioinformatics – the way from gene to drug" targeted for students (biophysics, biochemistry and medical cybernetics) on their 8th semester and special course "Bioinformatics and computer-aided drug design" for students on 11th semester of Medico-Biological Faculty of Russian State Medical University.

METHODOLOGY

Bioinformatics teaching is based on combination of traditional and modern approaches: giving lectures with using of computer presentations in PowerPoint program, seminars using direct access to the Internet bioinformatics resources, practical training and exercises at computer classroom, real scientific works in research laboratories, intermediate written tests and semester final examination.

Available hardware: SGI servers Origin-350 (8 CPU), 3 servers SGI Origin-200 (2 CPU), Linux cluster (32 CPU), 3 workstation SGI O2 and about 40 PCs.

Available software: molecular modeling suits Sybyl (Tripos Inc.), Amber, Gromacs, PASS, DockSearch, some academic programs, free software, etc.

IMPLEMENTATION

The main focus of our bioinformatics courses lies in the set of topics, which brief review is given below.

1. Bioinformatics: definition, purposes, tasks and place in modern biomedical science.
    - *Bioinformatics unites genomics and proteomics with classical disciplines (molecular biology, biochemistry, biophysics, molecular pharmacology).*
    - *Introduction in bioinformatics, the review of the basic purposes and problems.*
2. Integral platform "From gene to drug" *in silico* and *in vitro* (Ivanov, 2005).
    - *Bioinformatics technologies in the integrated platform "From genome to drug".*
    - *Search of new potential target proteins and experimental validation.*
    - *Experimental analysis and computer 3D modeling of target proteins.*
    - *Search and design new ligands based on 3D structure of target protein.*
    - *Prediction of lead compounds activity and in vitro testing.*
3. Basic principles of computer molecular modeling (Ivanov *et al*., 2005).
    - *Computational chemistry, methods and approaches.*
    - *Computer modeling of proteins 3D structures and their complexes.*
    - *Molecular mechanics (force fields, parameterization, energy minimization, conformation analysis, molecular dynamics simulation).*
    - *Quantum mechanics (ab initio and semi-empirical approaches).*
4. Analysis of biological texts.
    - *Software for storage, processing and analysis of biological texts.*
    - *Methods of sequence alignment of biopolymers.*
    - *Analysis of structure-functional similarity of new proteins.*
    - *Multiple alignment, patterns, motives and domains, proteins families.*
    - *Bioinformatics approaches in creation of new generation vaccines.*
5. Comparative genome analysis and search of new molecular targets.
    - *Comparative genomics and computer design of new antibacterial drugs.*
    - *Choice of new target in drug discovery pipeline.*
    - *Requirements to the new target proteins and criteria of choice.*
    - *Automated targets search: programs CATS, GenMesh.*

- *An example of new targets search for creation of new antituberculosis drugs.*
- *Further use of potential targets list.*
- *Protein function prediction at high and low homology with known proteins.*

6. Proteins superfamilies, Cytochrome P450 Database (CPD).
   - *Proteins superfamilies.*
   - *Proteins superfamily formation, structural similarity and functions variety of superfamily members.*
   - *Software for comparative analysis and classification of proteins in superfamily.*
   - *Cytochromes P450 – from superfamily formation to computer modeling and planning of genetic engineering experiments.*
7. Bioinformatics and Genomics.
   - *Tasks, problems and approaches of bioinformatics in genome data analysis.*
   - *Known full genomes: current situation, databases.*
   - *Full genome annotation.*
8. Bioinformatics and Proteomics.
   - *Proteins inventory.*
   - *Proteomics methods: 1D and 2D electrophoreses, HPLC, mass-spectrometry, bioinformatics.*
   - *Protein identification using proteins separation and mass-spectrometry analysis.*
   - *The modern bioinformatics approaches in applied proteomics.*
   - *Review of programs for the of 2D-electrophoresis image analysis, databases.*
   - *Algorithms of protein sequence identification based on peptides mass-spectra.*
9. Methods of analysis of protein 3D structures, database PDB.
   - *Methods of analysis of protein 3D structures, problems and prospects of structural molecular biology.*
   - *Proteins x-ray crystallography, problems of proteins crystallization.*
   - *NMR spectroscopy - physical bases, NMR analysis of protein structure.*
   - *Various types of microscopy.*
   - *Protein Data Bank (PDB).*
10. Computer 3D modeling of proteins. Cytochromes P450 modeling (Ivanov, 2002).
    - *Homology modeling of protein 3D structure.*
    - *Modeling of cytochromes P450* (Ivanov *et al*., 2003).
11. Proteins – molecular targets for drugs.
    - *Review of structures of known target proteins.*
    - *Searching of new targets.*
    - *Targets for creation of drugs against HIV.*
    - *Protein-protein interactions – new targets* (Veselovsky, 2002).
12. Computer technologies in drug discovery.
    - *Transformation of casual drug search into the rational drug design.*
    - *Computer methods in early stages of new drug search.*
    - *Review of computer technologies application in drug discovery.*
13. Structure-based drug design (Veselovsky, Ivanov, 2003).
    - *Role and place of computer drug design in classical scheme of drug creation.*
    - *Dogma of modern molecular pharmacology.*
    - *The basic stages of structure-based drug design.*
    - *Integration of computer and experimental methods in the way from gene to drug.*
14. Molecular docking, protein-ligand complexes, database mining.
    - *Algorithms of molecular docking.*
    - *Programs DOCK, AUTODOCK and LEAPFROG.*
    - *Using of molecular docking for modeling and lead search.*
    - *Program DockSearch, integration with Sybyl.*
    - *Searchig of cytochromes P450 ligands.*

15. Inhibitors of protein-protein interactions and protein dimerization.
    - *Strategy of design of protein-protein interaction inhibitors.*
    - *Inhibitors of HIV protease dimerization.*
16. Quantitative structure-property relationship (QSAR), lead search and optimization.
    - *Structure-activity relationship, QSAR.*
    - *Ligand-based drug design.*
    - *Known drugs designed with using computer technologies.*
17. *Prediction of biological activity spectra, creation of more safety and effective drugs.*
    - *Prediction of biological activity spectrum using compound structural formula.*
    - *Computer program PASS.*

DISCUSSION

Described bioinformatics courses can be useful for any students and post-graduates wishing to learn the bases of computer biochemistry, biophysics and pharmacology.

From 1995 this courses were passed by > 500 students (biochemists, biophysics, medical cybernetics) from Russian State Medical University, about 30 students from Sechenov Moscow Medical Academy (pharmacists), from Moscow Institute of Engineering and Physics (cybernetic), from Moscow Physics-technical Institute (biophysics), post-graduates from RAMS, as well as employees from the Institute of Biomedical Chemistry and the Institute of Pharmacology RAMS.

The best graduates of these courses made their diploma in the Institute of Biomedical Chemistry RAMS and later PhD dissertation as RAMS postgraduate.

REFERENCES

Ivanov A.S. (2005) Basic principles of computational chemistry for medical biologists. *Biomed Khim.*, **51**, 152–69. (In Russ.).

Ivanov A.S. *et al.* (2005) Bioinformatics Platform Development: From Gene to Lead Compound. In Larson R.S. (ed.), *Methods in Molecular Biology, vol. 316: Bioinformatics and Drug Discovery.* Humana Press, Totowa, pp. 389–432.

Ivanov A.S. (2002) Computer modelling of P450 structures. *FEBS Advanced Course 2002 "Cytochrome P450 Systems: from Structure to Application",* Kranjska Gora, Slovenia, 2002, 91–98.

Ivanov A.S. *et al.* (2003) General trends in 3D modelling of cytochromes P450. *Proceedings of 13th International Conference "Cytochromes P450, Biochemistry.* In Monduzzi (ed.), Biophysics and Drug Metabolism", Prague, D629R9045, 47–54.

Veselovsky A.V. *et al.* (2002) Protein-protein interactions: mechanisms and modification by drugs. *J. Mol. Recognit.*, **15**, 405–422.

Veselovsky A.V., Ivanov A.S. (2003) Strategy of Computer-Aided Drug Design. *Current Drug Targets - Infectious Disorders*, **3**, 33–40.

IN SEARCH OF GENETIC SIGNATURE FOR THE EXPANSION OF IRRIGATION SYSTEMS IN BALI

***Karafet T.M.*[*1, 2], *Lansing J.S.*[2], *Hammer M.F.*[2]**
[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia; [2] University of Arizona, Tucson, USA
[*] Corresponding author: e-mail: tkarafet@email.arizona.edu

Key words: genetic structure, haplogroup, haplotype, Balinese subaks, irrigation system

SUMMARY

Motivation: With the use of genetic markers it is possible to trace micro-migrations and changing patterns of relatedness within small communities in the recent past. We present a model to explain the origins and spread of irrigation systems on Balinese volcanoes. We then test some of the model's predictions by means of a comparative analysis of genetic relatedness in 21 villages and nine regions of Bali.

Results: The genetic analyses presented here provide several clear insights into the origins and expansion of the subaks, all of which are consistent with a kin-structured budding model of gradual expansion carried out by rice farmers.

INTRODUCTION

In Southeast Asia the spread of irrigated rice agriculture was usually linked with the expansion of precolonial kingdoms. Typically, the earliest irrigation systems were constructed by villages, and later consolidated and expanded by their rulers. But because of Bali's steep volcanic topography, the spatial distribution of Balinese irrigation canals made it impossible for irrigation to be handled at a purely community level (Christie, 1992). The problem was solved by the appearance of a new institution called subak. Subak were associations of farmers who shared water from a common source, such as a spring or irrigation canal. Very high level of cooperation within and between subaks has its roots in the ecology of the rice paddies, and the other in marital alliances (Lansing, 2005, 2006).

We can construct two alternative scenarios for the expansion of irrigation and rice cultivation, which would produce contrasting signals in the genetic structure of farming villages. If the expansion of irrigation was accomplished by the farmers rather than their rulers, then population movement of males would occur only as a result of demographic pressure leading to the formation of new daughter settlements close to the parent villages. This budding model would predict the formation of small communities located along irrigation systems, with the oldest settlements located at the irrigation outtakes located closest to the most ancient weirs or springs. Small population size and reproductive isolation would produce high rates of genetic drift; the older the community, the more evidence of drift. Patrilineages should exhibit less evidence of movement on the landscape than matrilineages, because only males inherit rights to farmland. In the alternative scenario in which the expansion of irrigation was managed by the state, none of these constraints would be in evidence. Overall, this scenario would predict a more fluid population structure with less nucleation of settlements and less genetic drift within settlements, compared with the budding model.

MATERIALS AND METHODS

A total of 507 Balinese males were analyzed in this study. One group consisted of 287 farmers from old 13 subaks in the vicinity of Gunung Kawi. The second group consists of 120 farmers belonging to eight younger and smaller subaks located in the regency of Tabanan. The third group comprises 100 random samples from each of the nine administrative districts on the island to provide context for the subak samples (Fig. 1). The Y chromosome polymorphic sites in our survey included 71 single nucleotide polymorphisms (SNPs) and 12 short tandem repeats (STRs). We also analyzed sequence data for the hyper-variable segment 1(HVS1-519 bp) of mtDNA. Parameters of within-population diversity, population genetic structure indices, and relationships between genetic and geographic structure by the Mantel test were estimated by using the ARLEQUIN 2.000 software (Schneider *et al*., 2000). The standardized measure G'_{ST} (Hedrick, 2005) was calculated to compare the levels of differentiation between Y-chromosome and mtDNA data. Median-joining (MJ) networks (Bandelt *et al*., 1999) were constructed by using the NETWORK 2.0c program.

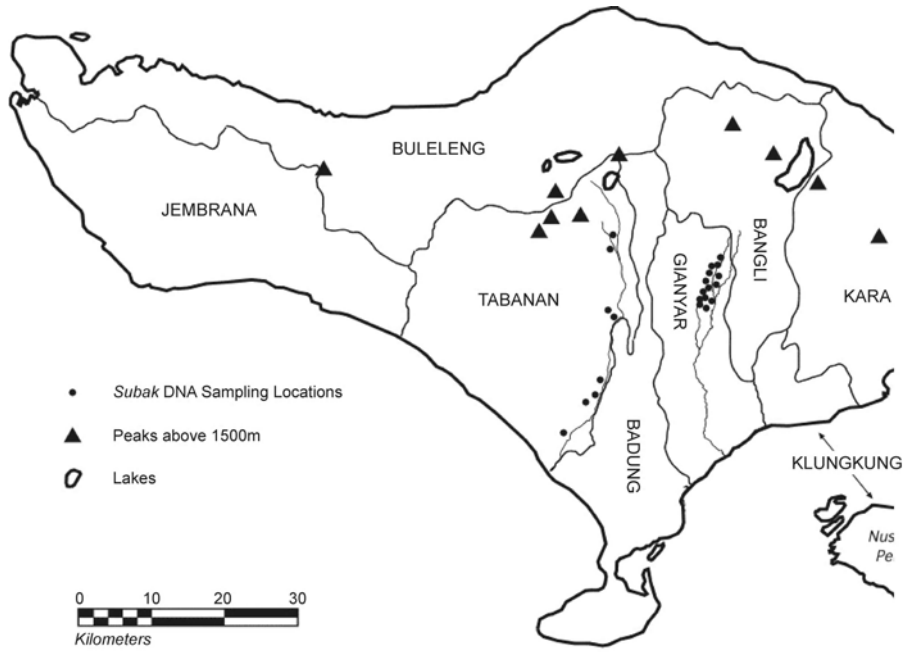


Figure 1. Map of Bali showing locations of Gunung Kawi (Gianyar) and Tabanan subaks, and boundaries of the 9 regions from which DNA samples were taken.

RESULTS

Genetic structure of Balinese subaks. The trend for all three genetic systems demonstrated lower diversity in the Gunung Kawi region, compared with Tabanan and Bali as a whole. Given the archaeological data indicating an older age of Gunung Kawi subaks compared with Tabanan, these results are consistent with a smaller effective population size for the Gunung Kawi region. This supports the hypothesis that Gunung Kawi subaks were established as small communities over a long period of time, perhaps experiencing sequential founder effects as predicted by the budding deme model. Moreover, since migrants comprised a biological kin group, migration may actually increase local genetic differentiation (Fix, 2004).

To further investigate the putative effects of serial founder effects on the Gunung Kawa subak system, we compared diversity parameters of the older and younger subaks of Gunung Kawi. As predicted by the budding deme model, the mean diversity in the younger Gunung Kawi subaks was lower than that in the older subaks. A budding deme model also predicts an increase in genetic differentiation among subpopulations as the process continues over time (Fix, 2002). The F_{ST} values for the Gunung Kawi subaks were notably higher for all three genetic systems than those for Tabanan subaks and the nine randomly collected samples in Bali, indicating a significant degree of genetic differentiation within Gunung Kawi region. For Y-SNPs Gunung Kawi exhibited a high F_{ST} value (0.141), nearly twice as high as for the regional Bali sample (0.075). We found higher F_{ST} value for mtDNA variation than for Y-STR variation in Gunung Kawi and Tabanan subaks.

Because the interpretation of genetic differentiation based on measures of F_{ST} is complicated by its dependence on levels of genetic diversity associated with different loci, we computed a standardized measure of genetic variation described by Hedrick (2005). This measure (G'_{ST}), which is related to the widely used G_{ST} parameter, allows more appropriate comparisons between loci with different mutation rates. G'_{ST} estimates showed similar levels of differentiation for mtDNA and Y-STR data in the Bali region (mtDNA G'_{ST}/Y-STR G'_{ST} = 1.054), while Gunung Kawi and Tabanan subaks demonstrated higher population structuring for Y chromosome (0.828 and 0.890, respectively). Moreover, within Gunung Kawi region Y chromosome differentiation was significantly higher for Y-SNPs and Y-STRs in older subaks (0.167 and 0.069, respectively) than in younger subaks (0.103 and 0.055, respectively), while mtDNA was less differentiated in older subaks (0.060) than younger subaks (0.072). These results are consistent with the predictions of the budding deme model: the level of endogamy is higher for males because of the strong tendency for patrilocality in rice-growing villages, while females migrate more often, albeit at only a low rate.

Genetic and geographic variation over short areas. To test for associations between genetic and geographic variation in Bali we performed Mantel tests. Geographic and genetic distances were not correlated for Y chromosome and mtDNA data in Tabanan or Bali as a whole. In contrast, Y-STR variation was highly correlated with geographic distances in Gunung Kawi (0.541), as was mtDNA variation (0.361). To investigate the correspondence between paternal and maternal genetic variation, we also estimated correlation coefficients among different loci. The results indicate a strong correlation between Y-STR and mtDNA structure in Gunung Kawi (r = 0.629), possibly reflecting the same events in population history. In contrast, there was no positive correlation among Y chromosome and mtDNA genetic distances in Tabanan or the geographic regions of Bali (r = –0.159 and r = –0.223, respectively).

DISCUSSION

The genetic analyses presented here provide several clear insights into the origins and expansion of the subaks, all of which are consistent with a kin-structured budding model of gradual expansion carried out by rice farmers. In this model, demographic pressure drives increases in population size, and downstream budding of the subak system. Subaks located at the furthest upstream positions on their respective irrigation systems demonstrate greater levels of genetic differentiation and diversity, suggesting that they must have been built before their downstream neighbors. The evidence from the Y chromosome is consistent with key features of the budding deme model: patrilocal residence with very little movement on the landscape except for occasional micro-movements to nearby daughter settlements. The older the subak, the more evidence for this pattern. Evidence from mtDNA is consistent with the contemporary observed pattern

of patrilocal residence and preferential village or subak endogamy, but occasional marriages outside the subak. Again, this pattern is most strongly evidenced in the oldest villages, and scales with time. The all-Bali sample shows none of these patterns.

We conclude with the observation that the budding model imposes a very restrictive set of constraints on the genetic structure of farming communities in Bali. These include strong founder effects accompanied by genetic drift and directional micro-movements; more structure in patrilineages than matrilineages, and a strong contrast between subaks versus background relatedness of the whole population. These are not the patterns expected under the alternative scenario of State controlled expansion of irrigation; that is, if the rajahs transported whole villages to newly constructed irrigation areas, or alternatively brought settlers from nearby villages.

ACKNOWLEDGEMENTS

This work was supported by the National Science Foundation grants BCS 0083524 and 043224 to JSL, MFH and TMK.

REFERENCES

Bandelt H.J. *et al*. (1999) Median-joining networks for inferring intraspecific phylogenies. *Mol. Biol. Evol.*, **16,** 37–48.

Christi J.W. (1992) Water from the Ancestors: irrigation in early Java and Bali. In Jonathan Rigg (ed.), *The Gift of Water: Water Management, Cosmology and the State in South East Asia.* London: School of Oriental and African Studies, University of London. pp. 7–25.

Fix A.G. (2002) Colonization models and initial genetic diversity in the Americas. *Hum. Biol.,***74**, 1–10.

Fix A.G. (2004) Kin-structured migration: causes and consequences. *Am. J. Hum.Biol.* 16, 387–394.

Hedrick P.W. (2005) A standardized genetic differentiation measure. *Evolution*, **59**, 1633–1638.

Lansing J.S. (2005) On Irrigation and the Balinese Stat'. *Curr. Anthropol.*, **46**, 305–306.

Lansing J.S. (2006) *Perfect Order: Recognizing Complexity in Bali.* Princeton University Press.

Schneider S. *et al*. (2000) *ARLEQUIN ver.2.000: A Software for Population Genetic Analysis*, Genetics and Biometry Laboratory, University of Geneva, Switzerland.

EPI-GIS: GIS ASSISTED COMPUTER TOOLS FOR DATA ACCUMULATION, COMPUTER ANALYSIS AND MODELING IN MOLECULAR EPIDEMIOLOGY

Kolchanov N.A.*[*1], Orlova G.V.[1], Bachinsky A.G.[2], Bazhan S.I.[2], Shvarts Ya.Sh.[3], Golomolzin V.V.[1], Popov D.Yu.[1], Efimov V.M.[5], Tololo I.V.[6], Ananko E.A.[1], Podkolodnaya O.A.[1], Il'ina E.N.[4], Rogov S.I.[4], Tretiakov V.E.[4], Kubanova A.A.[7], Govorun V.M.[4]

[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia; [2] The State Research Center of Virology and Biotechnology "Vector", Novosibirsk, 633159, Russia; [3] Institute of Internal Medicine, SB RAMS, Novosibirsk, Russia; [4] Institute of Physical and Chemical Medicine, MZ RF, Moscow, Russia; [5] Institute of Animal Taxonomy and Ecology, SB RAS, Novosibirsk, Russia; [6] Institute of Automation and Electrometry, SB RAS, Novosibirsk, Russia; [7] Central Research Institute of Dermatovenerology, MZ RF, Moscow, Russia

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

Key words: molecular epidemiology, GIS, tuberculosis, gonorrhoea, gene networks, mathematical modeling, antibiotic resistance

SUMMARY

Motivation: The system developed is aimed at accumulation, storage, retrieval, analysis, and visualization of experimental data in molecular epidemiology of socially significant diseases in Russia by applying mathematical modeling and Geographic Information Systems Technology. The combination of approaches is helpful in administrative decision-making, control of drug resistance spreading, and development of optimal strategy for decreasing epidemic level.

Results: The system includes the software tools for molecular-genetical monitoring of dynamics of distribution of both infectious disease and its causative agents, microorganisms. The system is supplied by the modules executing mathematical modeling of infectious diseases at the populational level and at the level of an individual. It contains the instruments for statistical data analysis and elaboration of optimal strategies for interventions lowering morbidity rate. The gene network technology is applied for searching for the key genes involved in occurrence of drug resistance.

INTRODUCTION

A pivotal task is a complex description and modeling of the infectious and epidemic processes at different hierarchical levels of organization of the biological system "pathogen-human host" including analysis and modeling of (1) molecular-genetical organization of pathogen (i.e., genome, proteome, and metabolome), (2) genotypical pathogen variants estimated by high-productive technologies (e.g., mass-spectroscopy); (3) gene networks controlling interaction of pathogen with the host organism; (4) molecular-genetical and evolutionary mechanisms of drug resistance; (5) dynamics of the infectious process and the mechanisms of its management at the level of particular organism; (7) features of epidemic process in populations and at the level of territories; and (8) the ways of management by epidemic process on the basis of optimal administrative decision-making. The task indicated is referred to the complex tasks of the

systems biology, which solution demands complex approach based on integration of modern bioinformational technologies, mathematical methods of modeling and data analysis. The EPI-GIS system developed by us enables (1) to make prognosis of distribution of pathogenic agents causing infectious diseases by accounting their genetical variability; social and populational structure of the risk groups of patients; ecological, geographical and etc. characteristics, which are important for preparation and management of preventive measures and choosing of the optimal anti-epidemic strategy including pharmacological aspects of the problem; (2) to model temporal and spatial dynamics of development of the infectious diseases in order to plan administrative anti-epidemic management aimed at optimization of costs needed for decreasing the morbidity level and preventing distribution of epidemic process. The discriminative feature of the EPI-GIS system is application of the geoinformational technologies (GIS). The current version of EPI-GIS is oriented for analysis of molecular epidemiology of two socially important pathogens: *M. tuberculosis* and *N. gonorrhoeae*, which are the causative agents of tuberculosis (TB) and gonorrhoea, respectively.

METHODS AND ALGORITHMS

Initial experimentally tested genetic data on molecular epidemiology are described, accumulated, and stored in specialized databases. Genotyping of *Micobacteria tuberculosis* (MBT) and *N. gonorrhoeae* was performed by high-technological mass-spectrometry. Antimicrobial susceptibility testing was determined in accordance with the standard methods. The genes (e.g., 17 genes of *N. gonorrhoeae*) were studied for the presence of mutations associated with antibiotic resistance. The primary experimental data are treated by means of originally developed software MMDMS (Molecular microbiology data management system) and then accumulated in the database "Infectious agent" stored in the Database executive system, which is realized via MS SQL Server 2000 (Fig. 1).

Gene networks. Reconstruction of gene networks describing the action of anti-TB and anti-gonorrhoea drugs antibiotic resistance and the mechanisms of multiple drug resistance was performed in accordance with the original technology GeneNet (Ananko *et al.*, 2005).

Modeling of individual pathogenesis. The TB infection model presented herein (section "Infection") is an extended modification of the mathematical model proposed by Marino S., Kirschner D.E. (2004) for the immune response induced by MBT in human lungs and lymph nodes. An initial-value problem for a set of differential equations describing the mathematical model of TB was addressed by using the method proposed by C.W. Gear (1971) for stiff differential equations with a variable order of approximation.

Epidemiology modeling. The program TBEPID_MOD is designed for modeling of TB epidemics in the closed population (see section "Epidemic" at Fig. 1). The following conditions are represented in the model: X, sensible to infection individuals; R, infected individuals, latent (immune); L_i, individuals infected by the i-th bacterial strain, but not eliminators of bacilli; T_i, individuals infected by the i-th strain and being eliminators of bacilli. By 0, we denote ordinary drug-sensitive bacterial strain. The program is realized via the difference scheme of the ordinary differential equations solutions. Initially, the model is based on the Perelman model supplemented by distribution into classes, where the patients are infected by the strains with different resistance to antibiotics ($i = 1, n; n <= 5$).

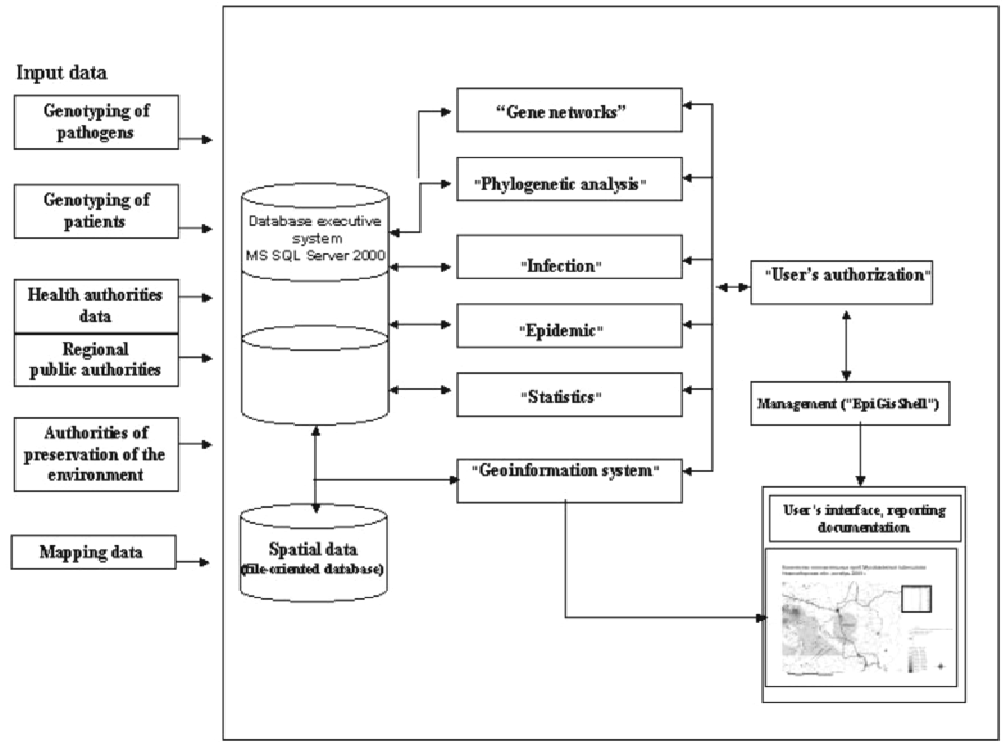


Figure 1. The structure of "EPI-GIS" system.

The model is based on the following system of equations:

$dX/dt = \pi X - \Sigma\beta_i(X)*(T_i+m_i*T0_i) - \mu X$

$dR_i/dt = \pi R_i + (rl_i+s_i)*L_j + (rl0_i+s0_i)*L0_j+(1-p)*\beta_i(X)*(T_i+m_i*T0_i) - (d_i*inf+\mu+\delta_i)*R_i$

$dL_i/dt = \pi L_i + (p_i*\beta_i(X)+rt_i+f_i+d_i*\Sigma\beta_i(R_j))*(T_i+m_i*T0_i) - (k_i+v_i+\mu+\mu l_i+rl_i)*L_i+\delta_i*R_i$

$dL0_i/dt = \pi L0_i + (rt0_i+f0_i)*T_i+k_i*L_i-(v0_i+\mu+\mu l_i+rl0_i+s_i)*L_i+\Sigma\gamma_{ji}*L_j$

$dT_i/dt = \pi T_i + v_i L_i - (kt_i+\mu+\mu t_i+rt_i+f_i)T_i$

$dT0_i/dt = \pi T0_i+v0_i*L0_i+kt_i*T_i-(\mu+\mu t_i+rt0_i+f_i)T0_i+\Sigma\eta_{ji}*T0_j dt.$

As a result, the intensity of selection of genetically modified pathogen forms could be estimated, as well as dependency of recovery rate upon the coverage and efficacy of treatment. Also, the program calculates the average year parameters: morbidity (novel cases per 100,000 of population), TB mortality (per 100,000 of population), and share of infected individuals in population. The computational module is realized via Delphi and operates under Windows or Linux.

Statistics. The block "Statistics" (including medical and social statistics) of EPI-GIS is developed for input data preprocessing and statistical analysis. The following statistical algorithms are realized in the system: (i) cutting out of drop-out data; (ii) calculation of statistical criteria such as mean value, standard deviation, coefficient of variation, data asymmetry, data excess, correspondence to normal distribution; (iii) calculation of correlation coefficient and linear regression; (iv) calculation of generalized integral characteristics by principal-factor method; (v) approximation of functional dependency between characteristics by linear regression or by neural network analysis.

GIS. GIS component is based on the product ArcGIS 9.0 developed by ESRI, the leading company in developing program software for GIS applications. The user-friendly platform ArcGIS corresponds to the standards enabling integration with other informational systems.

RESULTS AND DISCUSSION

The software of the EPI-GIS system operates via the distributed system. The server module contains the database server, web server, and the server component operating with geoinformational system ArcGis. The server module supports two variants of client applications: 1) client-oriented block, which is an extension of the product ArcMap entering ArcGis; and 2) window application for MSWindows2000/XP/2003 serving for data input/processing/querying and data visualization by the program ArcReader, which is freely distributed product of ESRI company.

The EPI-GIS system serves as an expert system supporting administrative decision making within the frames of the optimal strategy for medical treatment of an individual patient and control of anti-epidemic measures in conditions of increased transport flows between local populations and distribution of drug-resistant strains of microorganisms.

The general scheme of the EPI-GIS system is illustrated in the Fig. 1. The input data including statistical, medical, genetical, ecological, and geographical data are stored in the Database executive system. These data could be spatially visualized on a map by the tools of the "Geoinformation system".

GIS The system "EPI-GIS" contains the following main sections: "Gene networks", "Phylogenetic analysis", "Infection", "Epidemic", "Statistics", and "Geoinformation system". An access to these sections is provided through authorization of a user in accordance with the user's access rights. All these sections are integrated with the Management system ("EpiGisShell"), so that the data, where possible, could be visualized on a map and represented as reports in a form chosen by a user.

The gene networks of antibiotic resistance of mycobacteria and gonococci stored in the section "Gene networks" was reconstructed with the GeneNet technology (Ananko *et al.*, 2005). The subsection "*N. gonorrhoeae*: Resistance to antibiotics" of the GeneNet contains descriptions of 28 proteins, 24 genes, 16 small molecules, and 80 interrelations between components. This information has been extracted from 76 scientific papers. The subsection "Tuberculosis drug resistance" of the GeneNet contains description of 16 proteins, 15 genes, 10 small molecules, and 37 interrelations between components. The information has been extracted from 80 scientific papers. The gene networks technology enables to visualize the genes and proteins involved in different mechanisms causing resistance of pathogens to antibiotics. The section "Phylogenetic analysis" gives possibility to study relationships between the proteins influencing on drug-resistance in different species of microorganisms. The section "Infection" is designed for modeling of infection at the level of an individual patient. The software components of the section "Epidemic" were designed for modeling of epidemics by the example of TB at the level of population. As the model input, it is possible to order dynamical (distributed by time) interventions and migratory flows in calculation of dynamics of an epidemics. By mathematical modeling, it is possible to retrieve possible interventions and scenarios directed at improvement of epidemiological situation. The results of modeling obtained in the sections "Infection" and "Epidemic" may be statistically treated by the software of the section "Statistics" and visualized by the tools of the section "Geoinformation system".

Interventions. The EPI-GIS system contains a set of interventions that influence on epidemic or progression of individual pathogenesis. A user may choose them individually or in combinations. For example, the following interventions are considered in the section "Epidemic": (1) increase/decrease of the volume and efficacy of medical bed usage at specialized hospitals; (2) increase/decrease of the control/motivation for maintenance of the medical treatment prescriptions; (3) improvement of the treatment adequacy; (4) revaccination of healthy population, who lost postvaccinal immunity; (5) genotyping of DNA of both causative agent and patient, etc.

The extended model of individual pathogenesis described in the section "Infection" permits the simulation of TB treatment with one to five anti-TB drugs simultaneously and describes the emergence and further proliferation of genetically modified mycobacterium

strains with resistance to more than one anti-TB drug used under different treatment regimens. The extended model consists of ordinary differential equations containing 25 variables. Most coefficients and their value ranges were estimated from experimental literature data, the others, during the model verification. For more details, see Bazhan *et al*., this issue. The mathematical model presented allows predicting possible variants of disease courses and outcomes in individuals infected with MBT.

Using the model, treatments of TB infection under various different chemotherapeutic regimens (interventions) were imitated. It has been demonstrated that different regimes of the model, which correspond to different chemotherapeutic regimens, yield different outcomes, which is either recovery (adequate chemotherapy) or delayed recovery and the emergence of drug-resistant strains (inadequate chemotherapy).

Optimal administrative decision-making. This option of the EPI-GIS system is designed for computer-assisted decision-making aimed at the best distribution of administrative and financial resources directed on improvement of epidemic situation. This option is applicable for the sections "Epidemic" and "Infection". A hypothetic example of decision making for the section "Epidemic" is shown in Fig. 2. It illustrates epidemic prognosis of TB progression under condition that during 5 years local administration spends additional 5 mln rubles for anti-tuberculosis measures.

In Fig. 2, one may see the dependency of mortality from tuberculosis per 100,000 of population under different interventions and under optimal management by financial resources (0, no interventions; 1, increase of the number of hospital-beds; 2, increase of the therapy efficacy; 3, total revaccination; 4, genotyping of the pathogen; 5, genotyping of the patient; 6, enhancement of motivation for medical treatment; 7, optimal management by interventions). As seen, optimal management (curve 7) gives the best decrease of mortality progression.

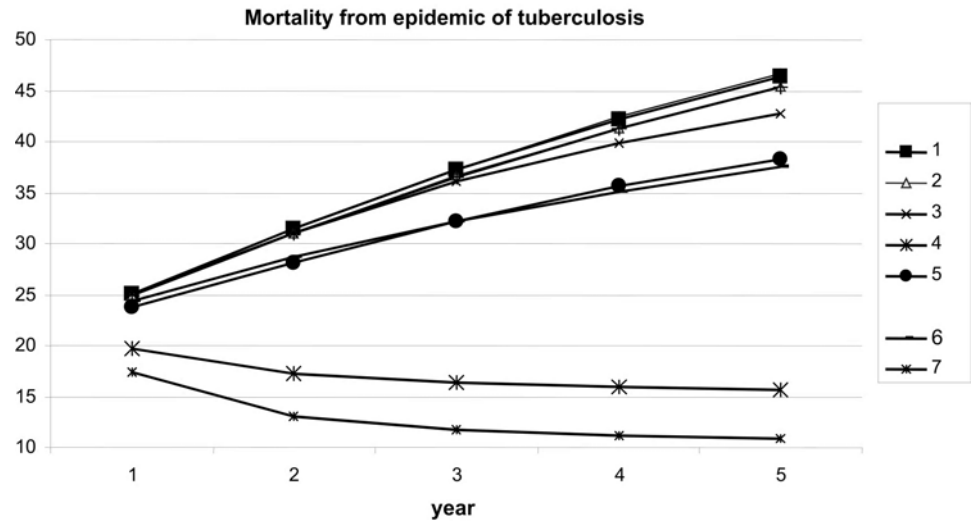


Figure 2. A hypothetic example of mortality from epidemic of tuberculosis.

Data output. The reporting documents may be obtained in a form of documents (Microsoft Office applications) via queries to the MS SQL server of the databases or by exporting bitmapped image from the window ArcGIS 8.1. The reports are generated via specialized menu in a form of plots, diagrams, and maps.

So, application of the system EPI-GIS may enhance targeted monitoring and control efforts, identification of risk factors, thus, providing the breakage of disease transmission and better recovery of patients.

ACKNOWLEDGEMENTS

The work was supported by the Russian Government Contract No. 02.434.11.3004.

REFERENCES

Ananko E.A., Podkolodny N.L., Stepanenko I.L., Podkolodnaya O.A., Rasskazov D.A., Miginsky D.S., Likhoshvai V.A., Ratushny A.V., Podkolodnaya N.N., Kolchanov N.A. (2005) GeneNet in 2005. *Nucl. Acids Res.*, **33**, D425–D427.

Bazhan S.I., Schwartz Ya.Sh., Gainova I.A., Ananko E.A. (2006) A mathematical model of immune response in infection induced by mycobacteria tuberculosis. prediction of the disease course and outcomes at different treatment regimens. *Proceedings of BGRS'2006. This issue.*

Gear C.W. (1971) The automatic integration of ordinary differential equations. *Communs. ACM*, **14**, 176–190.

Marino S., Kirschner D.E. (2004) The human immune response to *Mycobacterium tuberculosis* in lung and lymph node. *J. Theor. Biol.*, **227**(4), 463–486.

EVOLUTION OF THE STRUCTURE OF THE *XIST* LOCUS IN MAMMALS

Kolesnikov N.N.**, *Elisafenko E.A.*, *Zakian S.M.
Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia
* Corresponding author: e-mail: kolesnikov@bionet.nsc.ru

Key words: X chromosome inactivation, gene XIST, structural organization, evolution, mammalian XIST gene ancestor

SUMMARY

Motivation: The origin of sex chromosomes and the emergence of the mechanism of inactivation of the X chromosomes in the course of mammalian evolution remain some of the most intriguing phenomena. X inactivation is controlled by a special locus of the X chromosome: the inactivation center, in which the central functional role is assigned to the *Xist* gene. Despite progress, X inactivation is still in the spotlight and still promises surprises. The question of the origin of X inactivation is opened-ended: what molecular mechanisms made it happen? We are particularly interested in the question of the origin of the *Xist* gene. How and when did this gene emerge? On the basis of what sequences did they? And since this gene is in the class of large non-coding regulatory RNA genes, this question is more general. The aim of this work was to reconstruct the ancestral *Xist* gene of placental mammals, to reveal features of its structure and identify its evolutionary pathways.

Results: We have performed a comparative analysis of the structure and organization of this gene in a variety of evolutionarily close and distant mammalian species; revealed the mosaic evolution of the structure elements of the gene; identified common conserved and species-specific gene sequences; revealed an interspecific lability of the exon-intron structure; determined the structure of the gene for chimpanzee, dog and rat; reconstructed the gene, which was ancestral to *Xist* in placental mammals and about 100 Myr ago consisted of 10 exons and had most probably emerged de novo with the advent of the sex chromosomes.

INTRODUCTION

Inactivation of one of the two X chromosomes in female mammals is an impressive and large-scale example of epigenetic regulation, in the course of which most genes of one X chromosome in the females become switched off. It has been postulated that this phenomenon is associated with dose compensation in mammalian females (Lyon, 1961).

Transcriptional inactivation of one of the X chromosomes in mammalian females is a complex process. This process involves multiple chromatin modifications, which lead to the formation of stable facultative heterochromatin, and that heterochromatin is preserved in the succession of cell divisions.

This process is controlled by the inactivation center denoted as XIC (X – inactivation center; XIC, man; Xic, mouse), which spans about 1000 kb in the human X chromosome and about 450 kb in the mouse X chromosome. As has been demonstrated in recent years, *XIC/Xic* is a complex genetic locus, which contains important functional elements. Of them, the key element is the *XIST/Xist* gene (X-inactive specific transcript gene; *XIST*,

man; *Xist*, mouse). Transcription of this gene leads to the formation of non-coding RNAs spanning the entire chromosome and trigger a cascade of epigenetic chromatin modifications, which results in the inactivation *in cis* of the X chromosome (Borsani *et al.*, 1991; Brockdorff *et al.*, 1991; Brown *et al.*, 1991).

The *XIST/Xist* gene contains exons and introns and does not contain large ORFs. *XIST/Xist* RNA is subject to processing in the form of splicing into alternative transcripts and polyadenylation and stays in the nucleus. Regulation of this gene, in turn, is exerted by cis-elements and in mouse additionally by the antisense transcript of the *Xist* gene: the *Tsix* gene (Lee *et al.*, 1999). Human *TSIX* is reduced and functionally inactive compared to the mouse homologue, which suggests that regulation of the expression of the key gene must be different in different species (Migeon *et al.*, 2002).

Not only the *XIST/Xist* gene is involved in the process of inactivation, but also belongs to a so far small class of large non-coding regulatory RNA (ncRNA), the study of which has only recently been initiated and is believed to be promising (Furuno *et al.*, 2006).

MATERIALS AND METHODS

The common approach used consisted in analyzing the genomic sequences in the corresponding databases of sequenced genomes. The computer analysis was performed using the most recent versions of program packages such as BLAST (Altschul *et al.*, 1990, http://www.ncbi.nlm.nih.gov/), which searches for homologous sequences; TRF (Benson, 1999), for tandem repeats; IRF (Warburton *et al.*, 2004), for inverted repeats; RepeatMasker, (http://www.genome.washington.edu/UWGC/analysistools/repeatmask.htm><http://www.girinst.org), for mobile elements; Fasta (Pearson, Lipman, 1988) and CLUSTALX (Jeanmogin *et al.*, 1988), which align two or more sequences; the genomic analysis of large loci was performed using software programs and data available at servers (http://genome.ucsc.edu/; http://www.ensembl.org/) and the PipMaker software program (http://bio.cse.psu.edu).

The comparative analysis of the gene structure was performed on 10 species of four orders of placental mammals: *Rodentia, Cetartiodactyla, Carnivora, Primates*. We had previously determined the gene structure in four related vole species (Nesterova *et al.*, 2001), the gene orthologs in three species: chimpanzee, dog and rat, have now been identified by multiple and pairwise alignments with known human, mouse and bovine sequences (Chureau *et al.*, 2002).

RESULTS AND DISCUSSION

Gene structure. Overall, the structure of the *XIST/Xist* gene was found to be similar in all the species studied. However, of the 10 exons revealed, only six occur in all the species as part of mature RNA, and the others have species specific preferences. In doing so, we omitted the *Xist* RNA isoforms that resulted from alternative splicing from consideration. Thus, in man and mouse, this gene consists of 8 exons, of which only 6 are common to both. Exons 2 and 5 in man and chimpanzee are not what they are in rodents, but they are preserved in intron sequences in the inactive state (Table 1). We denoted the exons that are inactivated in some species but preserved in introns as pseudoexons (pEx). Three such pseudoexons, which are active in some species and inactivated in others, have been identified (Table 1). On the other hand, exon 2, which can be found in man, chimpanzee and dog, has been lost to the other species. The most interesting about it is that exon 2 expressed in man represents a small fragment of the 3′-region of L1MC3 transposon (Fig. 1), which is present in both chimpanzee and dog as part of the gene; however, as far as dog is concerned, this exon is inactive.

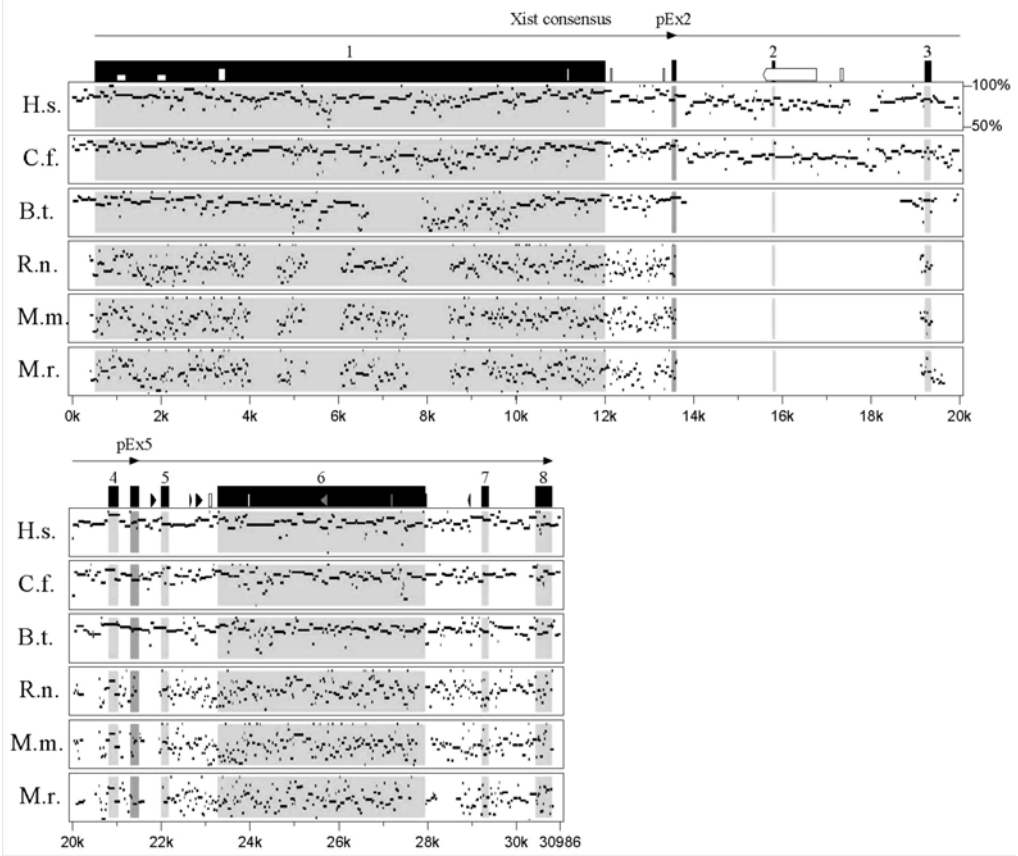


Figure 1. Comparative analysis of the *Xist* gene locus in placental mammalian species, the ancestral *Xist* gene versus genes in *Homo sapiens* (H.s.), *Canis familiaris* (C.f.), *Bos taurus* (B.t.), *Rattus norvegicus* (R.n.), *Mus musculus* (M.m.) and *Microtus rossiaemeriodianalis* (M.r.). Percent identity plot (PIP) of the ancestral *Xist* gene relative to various species is shown on the X axis, and the percentage of its identity (50–100 %) to various species is shown on the Y axis.

This is a striking example of how a transposon participates in the formation of the structure of a regulatory gene in man and chimpanzee: part of the transposon sequence is recruited as an exon.

Alongside the formation of pseudoexons, which arise due to mutation to intron-exon sites (pEx2h, man; pEx2p - chimpanzee), we could observe exons shortened at 3′ ends due to the formation of new 3′-5′ exon/intron junctions: Ex2r (rat), Ex7m (mouse), Ex7v (vole).

Thus, the *XIST/Xist* structure is not strictly conserved or stable from an evolutionary point of view: there is a lability of exon/intron transitions, which might be associated with species specific features of the gene functioning.

XIST/Xist gene ancestor. Pairwise and multiple alignment allowed us to reconstruct the consensus sequence of the ancestor gene, which existed in placental mammals about 100 Myr ago at the onset of species radiation, and to assess how far its structural elements have diverged from the consensus. The *Xist* ancestor gene consisted of 10 exons and its genomic locus was about 30 kb in size (Table 1, Fig. 1). As can be seen from the table, the overall length of the gene varies from 21 to 37 kb with a tendency towards a lesser size of the locus in *Rodentia* species (21–23 kb) and a greater size in the other species (32–37.5 kb). The shortening of the overall locus length in rodents is largely due to the shortening of intron lengths. The total length of exons is more than that of introns and the ratio of exon/intron length varies from 1.2 in man to 2.3 in cow, which makes a difference between the *Xist* gene and protein-encoding genes, in which the introns are ten to hundreds of times as long as the exons.

Table 1. Comparative size (in base pairs) of the *XIST/Xist* locus and its structural elements

Species	Gene size	Homologous exons										Total intron length	Total exon length
		Ex 1	**pEx2**	**Ex 2**	**Ex3**	**Ex 4**	**pEx5**	**Ex5**	**Ex6**	**Ex7**	**Ex 8**		
C.f.	37592	15480	96	59	140	211	162	130	4561	195	370	16188	21404
B.t.	34934	18693	90	–	137	210	164	131	4524	156	372	10457	24477
H.s.	32063	11333	90	64	137	209	161	164	4543	146	378	14838	17225
P.t.	32050	11316	90	64	137	209	144	164	4541	146	378	14861	17189
		Ex1	**Ex2**		**Ex3**	**Ex4**	**Ex5**	**Ex6**	**Ex7**	**pEx7**	**Ex8**		
R.n.	22898	9430	83	–	137	213	141	154	4610	146	387	7597	15301
M.m.	22786	9483	91	–	132	211	147	155	4521	141	340	7565	15221
M.r.	21161	7939	84	–	138	213	103	134	4361	134	384	7671	13490
anc.	30297	11506	93	58	138	209	185	169	4678	151	374	12736	17561
anc.		**Ex1**	**Ex2**	**Ex3**	**Ex4**	**Ex5**	**Ex6**	**Ex7**	**Ex8**	**Ex9**	**Ex10**		

Exons (Ex) and pseudoexons (pEx). Species: *Canis familiaris* (C.f.), *Bos taurus* (B.t.), *Homo sapiens* (H.s), *Pan troglodytes* (P.t.), *Rattus norvegicus* (R.n.), *Mus musculus* (M.m.), *Microtus rossiaemeriodianalis* (M.r.) and ancestor (anc); (–) – the exon is missing.

The most variable among all the exons was exon 1, for its length can vary twofold (cow - vole). This fact could be explained by the presence of tandem blocks of repeats (repeats A-F) within exon 1 (not shown). Some repeats (A, F, B, D) within exon 1, conserved and variously amplified, occur in all species; others are specific for a certain species (the C repeat for mouse, the B* repeat for man, the G repeat for cow). These repeats comprise 40 to 78 % of the entire sequence, as in the case of bovine ex-1, in which we have identified a new tandem repeat, 10 kb in length, which we denoted as the G repeat.

What additionally contribute to the variability of the exon 1 length are insertions of mobile elements, that is, SINEs, which, again, are species specific, except for MIR family members, which are common to all the mammals (Fig. 1).

All the other exons display slight variation in length (Table 1). The most conserved is exon 4 (Ex5a - ancestral), which contains a long inverted repeat; however, the function of this exon within gene RNA is not yet clear.

As early as when it existed, the ancestral gene contained the basic domains of tandem repeats, A-E, which suggests that their amplification occurred more that 100 Myr ago; whereas the amplification of the species specific tandem repeats and insertion of the species specific SINEs was confined to species radiation or earlier times. It is most likely that the presence of species specific sequences (exons, repeats, mobile elements) in the *Xist* structure reflects the way in which the gene function within the X chromosome to which it originally belongs, and apparently is a common feature to all the regulatory genes of that type.

Based on the presence of the remains of ancient mobile elements (L1MC3, L2, L3, MIR) in the exons and introns of the ancestral gene, their high level of divergence (over 30 %) and the results of comparison of the time of their active genome-wide spreading (about 200 Myr ago), we have concluded that the *Xist* gene might arise de novo during the first stage of mammalian sex chromosome formation: 240–320 Myr ago.

ACKNOWLEDGEMENTS

The work was supported in part by the Integration Projects of the Presidium of Russian Academy of Sciences (contract N10104-34/P-18/155-270/1105-06-001).

REFERENCES

Altschul S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215,** 403–410.

Benson G. (1999) Tandem repeats finder: A program to analyze DNA sequences. *Nucl. Acids Res.*, **27,** 573–580.

Borsani G. *et al.* (1991) Characterization of a murine gene expressed from the inactive X chromosome. *Nature*, **351,** 325–329.

Brockdorff N. *et al.* (1991) Conservation of position and exclusive expression of mouse *Xist* from the inactive X chromosome. *Nature*, **351,** 329–331.

Brown C.J. *et al.* (1991) A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome. *Nature*, **349,** 38–44.

Chureau C. *et al.* (2002) Comparative sequence analysis of the X-inactivation center region in mouse, human and bovine. *Genome Res.*, **12**, 894–908.

Furuno M. *et al.* (2006) Cluster of internally primed transcripts reveal novel long noncodimg RNAs. *PLoS Genetics*, **2**, 537–553.

Jeannmougin F. *et al.* (1998) Multiple sequence alignment with Clustal X. *Trends Biochem. Sci.*, **23**, 403–405.

Lee J.T. *et al.* (1999) *Tsix*, a geneantisense to *Xist* at the X-inactivation centre. *Nat. Genetics*, **21,** 400–404.

Lyon, M.F. (1961) Gene action in the X chromosome of the mouse (*Mus musculus* L.). *Nature*, **190,** 372–373.

Migeon B.R. *et al.* (2002) Species differences in *TSIX/Tsix* reveal the role of these genes in X-chromosome inactivation. *Am. J. Hum. Genet.*, **71**, 286–293.

Nesterova T.B. *et al.* (2001) Characterization of the genomic *Xist* locus in rodent reveals conservation of overall gene structure and tandem repeats but rapid evolution of unique sequence. *Genome Res.*, **11**, 833–849.

Pearson W.R., Lipman D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA*, **85,** 2444–2448.

Warburton P.E. *et al.* (2004) Inverted repeat structure of the human genome: the X-chromosome contains a preponderance of large, highly homologous inverted repeats that contain tested genes. *Genome Res.*, **14,** 1861–1869.

BioUML: VISUAL MODELING, AUTOMATED CODE GENERATION AND SIMULATION OF BIOLOGICAL SYSTEMS

***Kolpakov F.*[*1, 2], *Puzanov M.*[1, 2], *Koshukov A.*[1, 2]**
[1] Institute of Systems Biology OOO, Novosibirsk, 630090, Russia; [2] Design Technological Institute of Digital Techniques, SB RAS, Novosibirsk, Russia
[*] Corresponding author: e-mail: fedor@biouml.org

Key words: metamodel, simulation engine, Java, MATLAB, SBML, biological systems

SUMMARY

Motivation: Reconstruction of complex biological systems from a huge amount of experimental data requires a formal language that is suitable both for human and computer.

Results: BioUML is integrated extensible open source Java workbench that adopts visual modeling approach for formal description and simulation of complex biological systems. Its core is metamodel 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 SBML and CellML models can be expressed in terms of the metamodel. This formal description can be used both for visual depiction and editing of biological system structure and for automated code generation to simulate a model behavior. BioUML workbench provides two alternative simulation engines using Java and MATLAB: 1) Java simulation engine – workbench automatically generates and compiles Java code on the base of visual model (diagram) of a biological system. 2) MATLAB simulation engine – workbench automatically generates code for MATLAB and invokes MATLAB engine to simulate a model behavior. Both simulations engines passed 100 % SBML semantic test suite.

Availability: http://www.biouml.org.

INTRODUCTION

Reconstruction of complex biological systems from a huge amount of experimental data requires a formal language that can be easily understood both by human and computer. It is known that graphical depiction of complex system is the most suitable way of understanding of its structure by human. Graphical notation allows human to completely and formally specify model so computer programs can analyze the model and simulate its behavior (Lee, 2001). Thus the problem of modeling and simulating of complex systems can be significantly simplified for researchers by using computer systems providing visual modeling. This approach is widely used in engineering and computer science. Examples are: MATLAB/Simulink (http://www.mathworks.com), AnyLogic (http://www.xjtek.com), UML (http://www.omg.org/uml/).

BioUML (http://www.biouml.org) – Biological Universal Modeling Language – is integrated extensible open source Java workbench that adopts visual modeling approach for formal description and simulation of complex biological systems (Fig. 1). Another distinctive feature of BioUML workbench is tight integration with databases on biological

pathways, query engine allows user to find interacting components of the system and show results as an editable graph (Kolpakov, 2004).

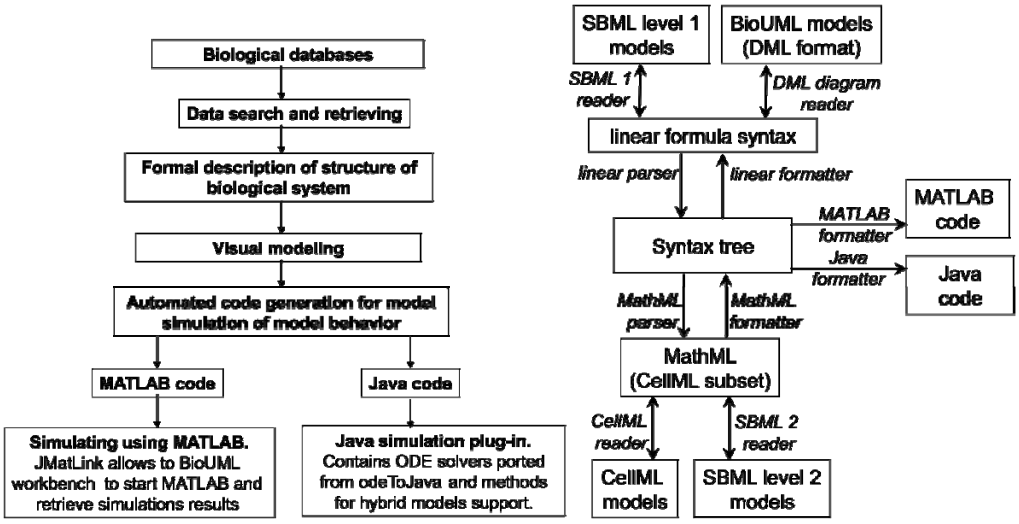


Figure 1. Data flow in BioUML workbench – left; parsing and conversions of mathematical expressions – right.

METAMODEL

The core of BioUML is a metamodel that provides an abstract layer (compartmentalized attributed graph) for comprehensive formal description of wide range of biological and other complex systems. Content of databases on biological pathways as well as SBML (Hucka *et al*., 2003) and CellML (Lloyd *et al*., 2004) models can be expressed in terms of the metamodel. This formal description can be used both for visual depiction and editing of biological system structure and for automated code generation to simulate a model behavior. Metamodel is problem domain neutral and splits the system description into three interconnected levels:

- graph structure – the system structure is described as compartmentalized graph;
- database level – each graph element can contain reference to some database object;
- mathematical model – any graph element can be element of mathematical) model.

Currently BioUML supports following mathematical elements: variable, formula, equation, event, state and transition. Fig. 2 demonstrates graphical depiction of these elements.

Special BioUML diagrams markup language (DML) is developed to store BioUML metamodel instance in XML format. Diagram structure description is divided into two parts: 1) diagram structure model – it describes the graph structure, location of diagram elements and contains references to associated with them database objects; 2) executable model – stores mathematical model associated with graph. Detailed description of DML format is available at http://www.biouml.org/dml.shtml.

SIMULATION ENGINE

Main parts of simulation engine are: code generator, formulas processor, algebraic equations solver and results writer. BioUML provides powerful formula processor that parses text and MathML expressions, result is presented as syntax tree and used by formatters to generate corresponding Java or Matlab code (Fig. 1, right).

Currently BioUML workbench provides two alternative simulation engines:

1) Java simulation engine – workbench automatically generates and compiles Java code on the base of visual model (diagram) of a biological system. For simulation we have adopted odeToJava library (Patterson, Spiteri, 2003) that provides methods for numerical solutions both stiff and non-stiff systems of ODEs. For solving algebraic equations Newton solver is used.

2) MATLAB simulation engine – workbench automatically generates code for MATLAB and invokes MATLAB engine to simulate a model behaviour using JMatlink library (http://www.held-mueller.de/JMatLink/).

Both simulations engines pass 100 % SBML semantic test suite that provides a set of valid SBML models with a simulated time course data (Finney, 2004). Test details are available at: http://www.biouml.org/sbml_tests/overview.html.

DISCUSSION AND FURTHER DEVELOPMENT

Several XML dialects like CellML, SBML are being developed for formal description and simulation of biological pathways. However they do not address problems of graphical notation for pathways visualization and tight integration with existing databases. Other approaches, for example BioPax (http://www.biopax.org), try to map information from different databases on biological pathways into common format; however BioPax format is not suitable for storage of mathematical models and simulation. The suggested approach should fill this gap – from one hand majority of models that can be expressed on SBML or CellML can be mapped into corresponding BioUML models, from the other hand information from different databases on biological pathways can be queried and presented as a set of diagrams.

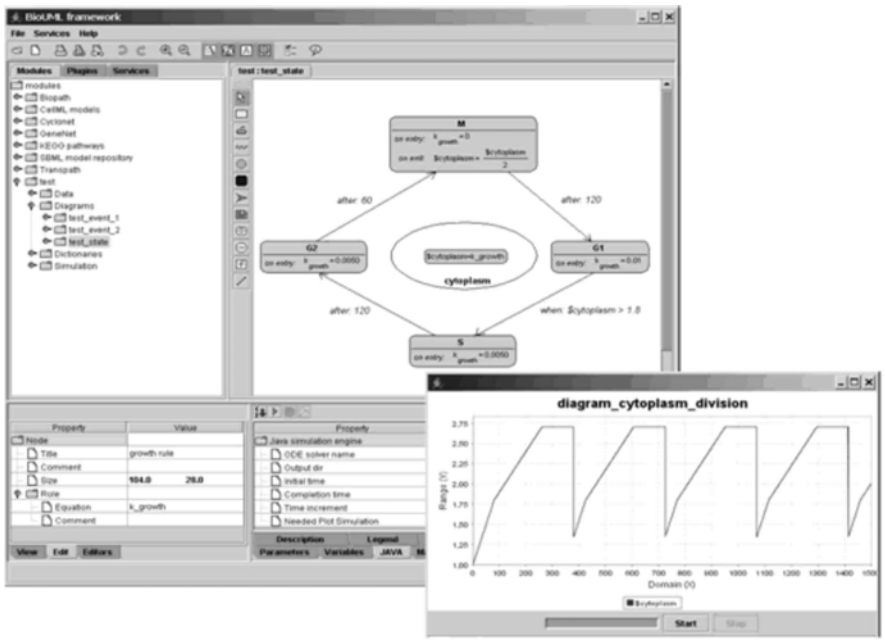


Figure 2. Simulation of toy hybrid model of cell cycle using BioUML workbench. Main cell cycle phases (G1, S, G2, M) are described as states. There are 4 transitions between these states: 3 transitions with time delay and 1 conditional transition – cell enters into S phase only when cytoplasm volume exceeds the specified threshold. Cell growth is described by simple ODE as rate of growth of cytoplasm volume, for different states growth rate is different.

Initial BioUML graphical notation and simulation engines have supported only ordinary differential equations for simulation of biological pathways (Kolpakov, 2004). Extended graphical notation (Fig. 2) and new version of simulation engines support piecewise functions, time delays, algebraic equations, events, states and transitions that allows user simulate wide range of biological systems. Nevertheless there are some limitations – mainly these are spatial models, for example models with diffusion and biomechanical models. We hope to overcome these limitations during further BioUML workbench development – now we are developing new plug-in for one-dimensional modeling of a vascular network in space-time variables. This task requires solving of one-dimensional partial differential equations (PDE).

Other direction of BioUML workbench development is connecting two worlds: world of cis-regulatory signals in DNA and world of molecular interaction networks of cells. For this purpose we are developing a set of plug-ins for analyses of nucleotide sequences and their visualization.

ACKNOWLEDGEMENTS

This work was supported by INTAS grant No. 03-51-5218, RFBR grant No. 04-04-49826-a and Siberian Branch of Russian Academy of Sciences (interdisciplinary projects No. 46).

REFERENCES

Finney A. (2004) http://sbml.org/wiki/Semantic_Test_Suite

Hucka M. *et al*. (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, **19**, 524–531.

Kolpakov F.A. (2004) BioUML – open source extensible workbench for systems biology. *Proceedings of BGRS'2004*, **2**, 77–80.

Lee E.A. (2001) Overview of the Ptolemy Project. Technical Memorandum UCB/ERL M01/11, University of California, Berkeley.

Lloyd C.M. *et al*. (2004) Progress in Biophysics and Molecular Biology, **85**, 433–450.

Patterson M., Spiteri R.J. (2003) http://www.netlib.org/ode/odeToJava.tgz

WEB SERVICES AT THE EUROPEAN BIOINFORMATICS INSTITUTE

Labarga A.*, *, Anderson M., Valentin F., Lopez R.
European Bioinformatics Institute, Hinxton, United Kingdom
*Corresponding author: e-mail: alabarga@ebi.ac.uk

Key words: web services, database integration, bioinformatics workflows

SUMMARY

Motivation: Following the exponentially growing amount of genomic sequence data from the different genome sequencing projects, and lately, of gene expression and protein interactions data, the challenge is now to unravel the gene functions, and to understand the gene regulation processes, and this genome-wide data analysis requires the complex interoperation of multiple databases and analytic tools. On the basis of these observations, the major world institutions for bioinformatics (like EBI, DDBJ, NCBI) have chosen to use the Web Services technology to expose its services in a programmatically accessible manner.

Results: We will show some of the tools we are developing a the European Bioinformatics Institute and how they can be used to construct analysis workflows using existing web services to help scientists to perform different tasks related to functional genomics experiments such as protein function prediction, microarray data annotation or literature data mining.

Availability: http://www.ebi.ac.uk/Tools/webservices.

INTRODUCTION

Today, biological databases are large collections of data that are relatively difficult to maintain outside the centers and institutions that produce them. These data and the corresponding analysis tools are mainly accessed using browser-based World Wide Web interfaces. When large amounts of data need to be retrieved and analyzed, this often proves to be tedious and impractical. Moreover, research is rarely completed just by retrieving or analyzing a particular nucleotide or protein sequence. Database information retrieval and analysis services have to be linked, so that, for example, search results from one database can be used as the base of a search in another, the results of which are then analyzed. When performing these operations using a Web browser, researchers are forced to repeat the troublesome tasks of searching; copying the results for subsequent searches to other databases, and again copying the results for analysis.

Creating a local bioinformatics work environment is possible by downloading and installing the necessary database content and services (such as retrieval and analysis programs). This has the advantage that processes that otherwise require manual operations can be automated. However, the amount of disk space required to store biological sequence databases can be huge, often exceeding several terabytes, requiring several hours, if not days, to complete analysis, even when using a supercomputer. For this reason, creation of a local system is not a suitable option for most individual researchers or institutions.

Web Services technology enables scientists to access biological data and analysis applications as if they were installed on their laboratory computers. Similarly, it enables programmers to build complex applications without the need to install and maintain the

databases and analysis tools and without having to take on the financial overheads that accompany these. Moreover, Web Services provide easier integration and interoperability between bioinformatics applications and the data they require.

METHODS AND ALGORITHMS

To ensure software from various sources work well together, this technology is built on open standards such as Simple Object Access Protocol (SOAP), a messaging protocol for transporting information; Web Services Description Language (WSDL), a standard method of describing Web Services and their capabilities, and Universal Description, Discovery, and Integration (UDDI), a platform-independent, XML-based registry for services. For the transport layer itself, Web Services can use most of the commonly available network protocols, especially Hypertext Transfer Protocol (HTTP).

A client (program) connecting to a web service can read the WSDL to determine what functions are available on the server. Any special data types used are embedded in the WSDL file in the form of an XML Schema. The client can then use SOAP to actually call one of the functions listed in the WSDL.

Services available

Currently, we support SOAP services for both database information retrieval and sequence analysis (Pilai *et al*., 2005). All available information about EBI web services can be accessed from the web page http://www.ebi.ac.uk/Tools/webservices.

Sequence and Literature data retrieval

WSDbfetch provides programmatic access to the popular sequence and literature data retrieval tool dbfetch (http://www.ebi.ac.uk/Tools/dbfetch). The databases currently available for data retrieval using this service include EMBL, EMBL-SVA, MEDLINE, UniProt, InterPro, PDB, RefSeq and HGVBase (Pilai *et al*., 2005).

CitationExplorer combines literature search with text mining tools for biology. It provides access to Medline, PubmedCentral, Patent Abstracts and Chinese Biological Abstracts databases. You can get full records from these databases, full text when available, and results are enriched with links to biological databases, synonyms, ontologies, etc.

OntologyLookup provides a web service interface to query any ontology available in the Open Biomedical Ontology (OBO) format.

Sequence analysis

The European Bioinformatics Institute also provides Web Services for sequence similarity tools (WSFasta, WSWUBlast, WSNCBIBlast, WSMPsrch, WSScanPS, WSScanWise); protein analysis (WSInterProScan); multiple alignment (WSClustalW, WSMuscle and WSTcoffee); and the European Molecular Biology Open Source Software Suite (WSEMBOSS), among others. These Web Services provide the same or even more advanced functionality than the traditional browser-based services described in (Harte *et al*., 2004).

RESULTS AND DISCUSSION

One of the main advantages of web services is that researches can construct easily bioinformatics workflows and pipelines combining two or more web services to solve complex biological tasks such as protein function prediction, genome annotation, microarray analysis, etc. These workflows can be created simple scripts, using advanced integration frameworks such as JBI, BPEL, etc or using scientific graphical workflow tools such as Taverna, developed at the EBI, or Triana, developed at the University of Cardiff.

Protein function prediction and classification

EBI web services are used internally to create new composite services such as InterProScan, (Quevillon *et al*., 1995) which combines different databases and protein signature recognition methods to provide automatic classification of proteins and function

annotation, or ProFunc (Laskowski *et al*., 2005) which has been developed to help identify the likely biochemical function of a protein from its three-dimensional structure using a combination of both sequence- and structure-based methods.

Gene expression analysis

In a typical microarray analysis workflow, the user might either upload gene expression data from an external source via the Expression Profiler (Kapushesky *et al*., 2004) web service, or retrieve data from the ArrayExpress public repository at the EBI (Labarga *et al*., 2005). The Data Selection method provides a basic statistical overview of the dataset, which can be used to guide the user in selecting genes relevant for further analysis. The Data Transformation methods can impute missing values in the chosen data subset, and perform other data transformations. Following these optional preprocessing steps, data can be subjected to one or more analyses. The user can explore the overall structure of the data via one of the clustering methods available in the Hierarchical and K-groups Clustering components, and the best number and quality of clusters within the data can be evaluated using a novel algorithm, Clustering Comparison. Alternatively, the user can subject data to a supervised method aimed at studying correspondences between groups of samples and genes in the Between Group Analysis component. Users interested in studying a specific gene or a group of genes can annotate them using WSDbfetch and map them to different ontologies using the OntologyLookup service, to get an insight on the molecular functions involved.

Genome annotation

For annotating a novel DNA sequence, users can use the WSGeneMark service to locate the predicted exons. Then they can use the EMBOSS tool sixpak to generate the 6-frame translation and perform a WSWUBlast search against Uniprot, get the homologous sequences with WSDbfetch and align them with WSClustalW. They can also submit the sequences to WSInterProScan for automatic identification of domains and classification.

CONCLUSION

Web Services technology brings into the bioinformatics community a new development concept in which users can access all data and applications hosted in the main research centers (EBI, DDBJ, KEGG, NCBI, etc) as if they were installed in their local machines, providing seamless integration between disparate services and allowing the construction of workflows to perform complex tasks.

REFERENCES

Harte N. *et al*. (2004) Public web-based services from the European Bioinformatics Institute. *Nucl. Acids Res*., **32**, 3–9.

Kapushesky M. *et al*. (2004) Expression Profiler: next generation—an online platform for analysis of microarray data. *Nucl. Acids Res*., **32**, Web Server issue.

Labarga A. *et al*. (2005) Web services at EBI EMBnet.news, **11**(4) 18–23.

Laskowski *et al*. (2005) ProFunc: a server for predicting protein function from 3D structure. *Nucl. Acids Res*., **33**, W89–W93.

Quevillon E. *et al*. (2005) InterProScan: protein domains identifier. *Nucl. Acids Res*., **33**, W116–W120.

EBI Web Services: http://www.ebi.ac.uk/Tools/webservices

EBI services: http://www.ebi.ac.uk/services

Profunc: http://www.ebi.ac.uk/thornton-srv/databases/ProFunc/

InterProScan: http://www.ebi.ac.uk/InterProScan

Expression Profiler: http://www.ebi.ac.uk/expressionprofiler

Taverna: http://taverna.sourceforge.net

Triana: http://www.trianacode.org

OBJECT-ORIENTED APPROACH TO BIOINFORMATICS SOFTWARE RESOURCES INTEGRATION

***Miginsky D.S.*[*1, 3], *Sokolov S.A.*[2], *Labuzhsky V.V.*[2], *Nikitin A.G.*[2, 3], *Tarancev I.G.*[4]**
[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia; [2] A.P. Ershov Institute of Informatics Systems, SB RAS, Novosibirsk, Russia; [3] Novosibirsk State University, Novosibirsk, 630090, Russia; [4] Institute of Automation and Electrometry, SB RAS, Novosibirsk, 630090, Russia
[*] Corresponding author: e-mail: shadow@bionet.nsc.ru

Key words: databases, resource integration, distributed software systems, object-oriented approach

SUMMARY

Motivation: Integration of different bioinformatics computational methods and databases is necessary to create an integrated "virtual laboratory" for specialists in genetics. The main purpose of such laboratory is to provide wide spectrum of modern computer methods of analysis of biological data and to apply such methods to data from different databases and other information resources all over the world from a single access-point. Such approach must severely increase the research efficiency in this and adjacent areas.

Results: Currently integration technologies for computational methods and databases are developed and prototyped. Both technologies based upon and cooperated by general set of linked terms representing biological concepts.

INTRODUCTION

A large amount of different bioinformatics software is available at the moment: different databases, computational analysis methods, presentation and visual modeling tools. When solving a complex problem one could find appropriate software for solving almost any small part of it. But in practice, if number of parts is more than three or four, assembling them into the single integrated tool capable to solve the proble is nearly impossible because of incompatibilities of different software programs. Programs could have different input/output formats and user interfaces, availability of source code etc. Using commercial software in most cases is not a way out either. There are some integrated solutions for some complex problems that one could buy, but there is almost zero probability to find such solution for a particular problem.

One of the ways to solve this problem is to integrate different programs and databases into the single standardized infrastructure that provides its' compatibility and combine them for current problems. We've found the way to develop the infrastructure using object-oriented approach. The general idea is that it's possible to find a set of terms (data types, classes) representing the concepts of genetics and unifying the data operated by concrete programs and databases in spite of incompatibility. By this way we are able to reduce the problem of incompatibility between programs to the problem of incompatibility between formats. Then allowing each term to have more than one representation (i.e. format) and providing converters between different formats of the same term we can totally eliminate such incompatibility. The software infrastructure that could manipulate and automatically convert data, provide data persistence and manage all integrated software is required in this case.

METHODS AND ALGORITHMS

An object-oriented approach appliance to the integration problem consists of two key integration efforts: integration of data and integration of behavior. The integration of data effort unifies disassembled knowledge clusters in the biology fields and provides the information ground for further research. The integration of behavior allows for reusing algorithms and software components in different appliances and provides a single base tool set for researchers.

Both integration solutions form the single workspace where programs could access databases by making queries and storing data.

DATA INTEGRATION SOLUTION

The set of resources could be seen as a heterogeneous network of interconnected databases containing knowledge on different biology (genetics first of all) fields. Connections between databases are represented either as direct links (when a database contains object identifiers from the other one), or as indirect links by matching set of properties (e.g. matching by name). Particular database is characterized by its model (terms and relations it uses) and authority area (particular field, such as gene expression regulation).

Current sizes of database records vary from several thousands of bytes to tens of megabytes, so it is hardly possible to fully download all the data from the network and store it in a single place. Fortunately, each data-object could be divided into 2 parts: relatively lightweight index data (for search purposes) and raw data (for computer analysis purposes).

The integration of data can be performed with using of the lightweight first parts (index data) only. The solution consists of a global index database (IDB) containing all index parts of data-objects from each integrated database, and drivers – one for each database – for database-specific access, i.e. extracting index and raw information, updating information, etc. The IDB contains enough information to find particular objects or set of objects by single query and it's doesn't matter in which database it is contained.

In order to integrate index data from different databases into single information space, it is necessary to develop specific IDB model that is flexible enough to contain relatively heterogeneous information on the one hand, and structured enough to effectively execute searching queries on the other hand. We developed an object-oriented meta-model that satisfies both requirements. Prefix "meta" means that there is no concrete database tables or other entities representing concrete biological concepts but one can describe any concrete set of interconnected concepts representing for example genetics field using meta-model.

The meta-model consists of the following key concepts (Fig. 1):

Term. Each term describes class of objects with common properties and relations to the other objects. The examples of terms are concrete biological concepts: "Gene", "Species", "Reaction".

Property. Each term is not simply the name but has internal structure represented by properties. Property is characterized by its data type, multiplicity and constraints. Properties could be either scalar or vector i.e. representing relationships between instances of terms, association-kind relationships first of all.

Having the particular model – a set of terms interconnected by their relation properties – one can integrate databases by mapping their structure on the model described. The mapping could be incomplete i.e. only some fields that are necessary for searching are mapped.

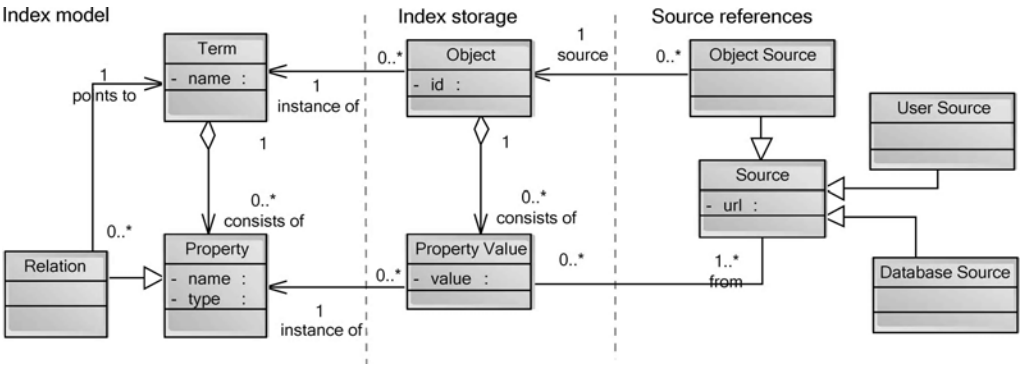


Figure 1. Integration database conceptual model.

While IDB imports the information on the per database level, it is also possible to integrate objects from different databases into single object in the IDB. Equivalency of two objects of the same term is matched by expert-defined criterion. If the criterion is met, two objects are merged automatically, while keeping full information about the sources of data (represented by the *Source* concept in meta-model). This key feature makes it possible to supplement information from one database by other one, compare them, look for variances and errors. It also guarantees that merge operation is reversible – information from one database could be split back from another one if necessary.

The integrated data-space provided by IDB is accessible by either object Java-API or specially-designed SQL-like query language – Index Query Language (IQL). IQL operates with terms and properties described. For example one could execute the following query within IDB instance with "Gene" and "Species" terms interconnected by gene's property "organism":

```
SELECT Gene WHERE (Gene.)name = "Apo B" AND
(Gene.)ogranism.latinName = "Homo sapience"
```

The results all this query will be the data about this particular gene (single gene because it's exactly identified in query) stored in all integrated databases.

BEHAVIOR INTEGRATION SOLUTION

Behavior of the data is represented by a set of different programs operating with them – methods. Each method in our model is characterized by a set of input and output data channels (*Pins*) and set of tuning parameters. Each input/output pin is characterized by particular term representing a kind of data it could obtain or produce. The set of terms is the same as used for databases integration purposes. A specific wrapper must be developed for each term in order to it to be supported by the methods – this is the expenses of integration. It is possible to use data from the IDB to feed these methods and store results of their execution back into the IDB.

Methods are subdivided into interactive and non-interactive ones. Non-interactive methods use input data and tuning parameters to produce data. Such methods have no user interface and could be viewed as a kind of tunable filters. Such filters could be combined by their inputs and outputs into scenarios to solve complex problems with multiple atomic steps. Methods formalization based upon terms guarantees that methods with pins conformed by terms are compatible. Thus, one is able to construct scenarios of any complexity based on existing methods and even other scenarios. The dataflow model is used. Scenarios are executed as a single complex method by specially designed virtual machine (VM). VM is able to run

scenarios with appropriate functionality of methods' wrappers in the distributed environment. For example, some methods could be performed by a local computer and others by a high-performance cluster. It is possible to run scenarios in the foreground (monitoring the process in real time), in the background, suspend and hibernate them.

Interactive methods are used in scenarios as an entry or exit points, and in most cases they are usable themselves. Such methods are usually designed for either creating or editing data by user or for data visualization.

RESULTS AND DISCUSSION

Consequently, we have an object-oriented approach to resources integration on bioinformatics field based on a set of terms (object classes) where each term is defined as the data representation form (object attributes) and as a set of methods to manipulate its instances (object methods). Scenarios are used as a high level programming language and allow specialists in genetics to perform scientific research *in silico*.

Currently the IDB is prototyped and two databases – GeneNet (Ananko *et al*., 2005) and SWISS-PROT (Boeckmann *et al*., 2003) – are integrated for testing purposes. KEGG (Kanehisa *et al*., 2006) and TRRD (Kolchanov *et al*., 2002) are currently being integrated. Also we are planning to build new generation of GeneNet database to entirely base on this technology that is more closely considered in work (Miginsky *et al*., 2006). IQL semantics is currently specified but syntax and some additional features are under development. VM is in alpha-version state and being tested with different scenarios.

ACKNOWLEDGEMENTS

The work is supported by the innovation project of Federal Agency of Science and Innovations IT-CP.5/001 "Development of software for computer modeling and design in postgenomic system biology (system biology *in silico*)".

REFERENCES

Ananko E.A., Podkolodny N.L., Stepanenko I.L., Podkolodnaya O.A., Rasskazov D.A., Miginsky D.S., Likhoshvai V.A., Ratushny A.V., Podkolodnaya N.N., Kolchanov N.A. (2005) GeneNet in 2005. *Nucl. Acids Res*., **33**, D425–D427.

Boeckmann B., Bairoch A., Apweiler R., Blatter M.-C., Estreicher A., Gasteiger E., Martin M.J., Michoud K., O'Donovan C., Phan I., Pilbout S., Schneider M. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucl. Acids Res*., **31**, 365–370.

Kanehisa M., Goto S., Hattori M., Aoki-Kinoshita K.F., Itoh M., Kawashima S., Katayama T., Araki M., Hirakawa M. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucl. Acids Res.*, **34** (Database issue):D354–357.

Kolchanov N.A., Ignatieva E.V., Ananko E.A., Podkolodnaya O.A., Stepanenko I.L., Merkulova T.I., Pozdnyakov M.A., Podkolodny N.L., Naumochkin A.N., Romashchenko A.G. (2002) Transcription regulatory regions database (TRRD): its status in 2002. *Nucl. Acids Res.*, **30**, 312–317.

Miginsky D.S., Suslov V.V., Rasskazov D.A., Podkolodny N.L., Kolchanov N.A. (2006) Architecture of software toolkit for storing and operating with biosystems models. *This issue.*

ARCHITECTURE OF SOFTWARE TOOLKIT FOR STORING AND OPERATING WITH BIOSYSTEMS MODELS

***Miginsky D.S.*[*1, 2], *Suslov V.V.*[1], *Rasskazov D.A.*[1, 2], *Podkolodny N.L.*[1], *Kolchanov N.A.*[1]**
[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia; [2] Novosibirsk State University, Novosibirsk, 630090, Russia
[*] Corresponding author: e-mail: shadow@bionet.nsc.ru

Key words: ecosystems, gene networks, visual modeling, databases

SUMMARY

Motivation: Each biosystem – from cell-level to biosphere – could be viewed as a self-regulating hierarchical system of elements interconnected into the network by the regulatory circuits. On different stages of biosystem's evolution such a network could contain various number of elements and has various structure of interconnections that is the consequence of serious behaviour variations. Regarding to the biosystem's type there are different types of elements and interconnections forming the domain of current biosystem (Zherikhin, 1994; Suslov *et al*., 2006). For instance the cell models contains microbodies, proteins and chemical substances as elements while the ecosystems includes concrete organisms, other biotic and abiotic factors, food (Luczkovich *et al*., 2003) and also completely different interaction types (De Jong-Brink, 1995; Schultze, Kondorosi, 1998; Adamo, 2002). The software capable to operate with different network types first of all need to be flexible enough and have an extensible domain. More precisely, each network type is described by separate domain. For each one user need to define the appropriate element types – and this must be the key feature of the software to develop.

Results: Currently the database capable to store different types of network is prototyped. The first two network types to be stored are gene and ecological networks. The database is capable to represent them, describe and associate with others stages of biosystem's evolution in context of single model and also decompose such networks.

INTRODUCTION

The value of data on any biological system is greatly increasing if formalized and represented in structured computer database. Such a representation provides wide specter of capabilities to analyze it, share between many experts all over the world and integrate with other information resources of the same domain. Currently most databases on this field have different disadvantages: (i) have very complex multilayered structure with inconvenient navigation or (ii) their structure too inflexible to extend the domain (set of concepts) or (iii) not structured and formalized enough, e.g. represented with text unstructured files (Sergeyev et al., 2006). For small-scale models of tissues, organs and also gene and protein networks there are also some schemas (iv) containing elementary objects and interconnections with or without spatial location information – a sort of object-oriented models (Ananko et al., 2005). Domains of such biosystems are more clearly understandable and could be described as a finite set of well-defined concepts and interconnections between them. The problem to solve is to extrapolate such an object-oriented (OO) approach to large-scale systems such as ecosystems.

In this work we tried to use the experience of operating with cell-level models – gene networks (Ananko et al., 2005). The software operating with networks of any level have to operate with different domains. Thus there must be the capability for user to define new domains by himself. Next he can build his own networks based on concepts described in chosen domain. If necessary this database could be integrated or supplied with information from other databases with any of i-iv structures.

Also there must be relatively simple and effective software capable to edit such networks. The most natural and effective solution is a sort of visual modeling software tool using vector 2D-graphics to represent networks as marked graphs with defined layout. This solution also gives the capabilities to decompose the network. For example user could hide some elements of network or visualize some functional aspects only while hiding others.

METHODS AND ALGORITHMS

Regarding to the requirements described earlier the software toolkit need to consist of two main subsystems – database and visual editor. The most important is the first subsystem because it defines the most part of business-logic i.e. all concepts operated by the software and all operations with them.

It is better to think about this database as high-level specific domain-oriented DBMS (databases management software) with its own meta-model (Fig. 1). User could define his own "database structure" based on this meta-model. By the database-structure here we mean the specific *Domain* that contains a set of *Terms* (it could be think as OO class) each one representing specific concept from the domain. Each term has a name (for gene networks it could be "gene" or "protein" for instance) should be defined as a set *Properties* characterizing it. For instance term "gene" characterized by its "name", list of alternative names ("synonyms"), link to "species" where it presented and link to "protein" it produces. Here each quoted name is a name of concrete property. Properties could be of different types and also could be either scalar or pointing (links). Links are very important because represents the relationships (associative first of all) between different concepts. Each link property is characterized by a set of terms to point from the same domain. Here is the first level of meta-model.

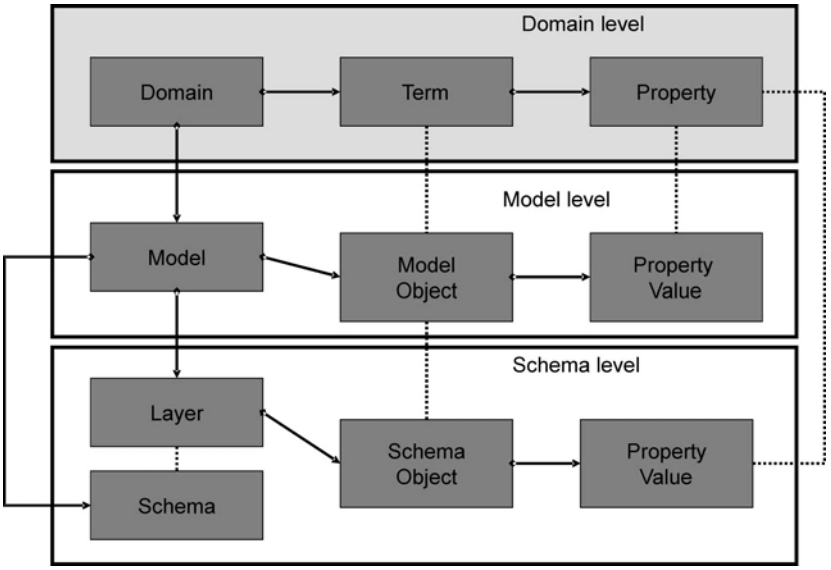


Figure 1. Database meta-model.

The second and third levels defines the concrete *Objects* as instances of terms. All objects are grouped into *Models* – a sort of logical partitions treated to concrete domain. For user the model is the representation of concrete biological system. Each object is contains values of properties defined by its term.

The most part of information on each individual object is described on the second level. It could be viewed as a sort of dictionary containing elements for networks. Almost all properties gain the values on this level.

The third layer has almost the same structure as the second but assigned for representation of networks. The entire model with its objects of the 2nd and 3rd levels could be viewed as a network representing concrete biological system. Third level contains *Schemas* representing some parts or functional aspects of the network. Its combination describes the entire network. For smaller network there could be the only one schema but it is hardly possible to find so simple biosystem that could be described in that way. So schemas are one of the ways to decompose the model into smaller parts. There are two other mechanisms of decomposing – *Layers* and compartments. Layer is a reusable part of the schema – more than one schema could contain the same layer. Layers are also necessary for functional decomposition so each schema contains the set of depended ones. Viewing the schema user can show or hide some of them. Creating or editing one he could reuse already created layers.

Compartments unlike layers and schemas provide spatial decomposition. Any object could be a compartment if its term is marked as compartment. For instance the "cell nucleus" term could be the compartment in gene networks. In meta-model the compartment is defined as schema level object able to contain others. The hierarchy of compartments could be of any nesting level. Viewing the schema user could show or hide payload of any compartment and define the necessary detail level by this way.

RESULTS AND DISCUSSION

Currently the database has been prototyped. Implementation of this model is based on the general database integration technology (Miginsky *et al*., 2006). This meta-models are pretty close in both works regardless to different purposes but this one has one additional level. This problem could be solved by mapping the domain into two separate but correlated domains from that one. For testing purposes we have imported some schemas from GeneNet (Ananko *et al*., 2005) into this database. The functional prototype of the editor is implemented in GeneNet. Currently new version of the editor oriented for described meta-model is being developed.

The first two areas of application of this software are gene and ecological networks. For gene networks it counts as next generation of GeneNet system. In comparison with it this one provides more flexible mechanisms of operating with this networks by new methods of decomposition and plug-in architecture capable to integrate modules for analysis.

ACKNOWLEDGEMENTS

Authors are grateful to prof. M.G. Sergeev and dr. N.I. Yurlov for assistance in results discussion. The work is supported by the following grants: RFFI 03-04-48506, integration projects SB RAS 34, project "Computer Modeling and Experimental Design of Gene Networks", RAS program on physicochemical biology, RAS presidium program "Biosphere Origin and Evolution" and innovation project of Federal Agency of Science and Innovations IT-CP.5/001 "Development of software for computer modeling and design in postgenomic system biology (system biology *in silico*)".

REFERENCES

Adamo S.A. (2002) Modulating the modulators: parasites, neuromodulators and host behavioral change. *Brain Behav. Evol*., **60(**6), 370–377.

Ananko E.A., Podkolodny N.L., Stepanenko I.L., Podkolodnaya O.A., Rasskazov D.A., Miginsky D.S., Likhoshvai V.A., Ratushny A.V., Podkolodnaya N.N., Kolchanov N.A. (2005) GeneNet in 2005. *Nucl. Acids Res*., **33,** D425–D427.

De Jong-Brink M. (1995) How schistosomes profit from the stress responses they elicit in their hosts. *Adv. Parasitol.*, **35,** 177–256.

Luczkovich J.J., Borgatti S.P., Johnson J.C., Everett M.G. (2003) Defining and measuring trophic role similarity in food webs using regular equivalence. *J. Theor. Biol.*, **220(**3), 303–321.

Miginsky D.S., Sokolov S.A., Labuzhsky V.V., Nikitin A.G. (2006) Object-oriented approach to bioinformatics software resources integration. *This issue.*

Schultze M., Kondorosi A. (1998) Regulation of symbiotic root nodule development. *Annu. Rev. Genet.*, **32,** 33–57.

Sergeyev M.G., Suslov V.V., Miginsky D.S. Yurlova N.I., Kolchanov N.A. (2006) Towards a database for ecosystem description using network technology. In *Ecosystem biodiversity and dynamics: informatics-based approaches and modeling*. SB of RAS Publisher, Novosibirsk, in press. (in Russ.).

Suslov V.V., Sergeev M.G., Yurlova N.I., Miginsky D.S. (2006) The ontology of ecosystems. *This issue.*

Zherikhin V.V. (1994) Evolutionary Biocenology. The Choice of Models. In *Ecosystem Rearrangements and Biosphere Evolution.* Nedra, Moscow, pp. 13–20.

AUTOMATED TEXT ANALYSIS OF BIOMEDICAL ABSTRACTS APPLIED TO THE EXTRACTION OF SIGNALING PATHWAYS INVOLVED IN PLANT COLD-ADAPTATION

Olsson B.*[*]*, Gawronska B., Erlendsson B., Lindlöf A., Dura E.
School of Humanities and Informatics, University of Skövde, Sweden
[*] Corresponding author: e-mail: bjorn.olsson@his.se

Key words: text analysis, information extraction, pathways, computational linguistics

SUMMARY

Motivation: Automated text analysis is an important tool for facilitating the extraction of knowledge from biomedical abstracts, thereby enabling researchers to build pathway models that integrate and summarize information from a large number of sources. Advanced methods of in-depth analysis of texts using grammar-based approaches developed within the field of computational linguistics must be adapted to the special requirements and challenges posed by biomedical texts, so that these methods can be made available to the bioinformatics and computational biology communities.

Results: Our system for automated text analysis and extraction of pathway information is here applied to a set of PubMed abstracts concerning the CBF signaling pathway, which is a key pathway involved in the cold-adaptation response of plants subjected to cold non-freezing temperatures. The system successfully and accurately re-discovers the main features of this pathway, while also pointing to interesting and plausible new hypotheses. The evaluation also reveals a number of issues which will be important targets in the continued development of the system, e.g. the need for an extended lexicon of taxonomic terms and an improved procedure for recognition of sentence boundaries.

INTRODUCTION

We have proposed a grammar-based method for extraction of biological relations from scientific texts (Gawronska *et al*., 2004; 2005b), which uses an algorithm to search through syntactic trees produced by a linguistic parser, identify relations mentioned in the sentences, and classify them with respect to semantic class and epistemic status (facts, counterfactuals, hypotheses). The semantic categories used for classification are based on the KEGG relation set (Kanehisa *et al.*, 2004), so that pathway maps in that notational convention can be automatically generated. We recently presented extensions of the method, such as improved named entity recognition and the addition of an algorithm for distinguishing between text describing previous and current work, thus making it possible to avoid extracting relations from those parts of the text which merely report findings from previous work or common knowledge (Gawronska *et al.*, 2005a; 2005b; Olsson *et al.*, in press). The syntactic analysis and the algorithm that derives pathways from syntactic trees are here only shown in a high-level overview (Fig. 1; for more detailed information, see (Gawronska *et al.*, 2005b) and for an overview of other text mining approaches for biology, see for example (Hirschman *et al.*, 2002)). Here, we apply the system to a set of PubMed abstracts concerning the CBF signaling pathway – a key

pathway involved in the cold-adaptation response of plants. It is of particular interest since the CBF/DREB1 transcription factors have been shown to be regulators of the majority of genes responsible for plant cold-acclimation (Bräutigam *et al.*, 2005; Thomashow, 1999), thus making them potential targets for efforts to genetically engineer crops with increased cold hardiness (Bräutigam *et al.*, 2006).

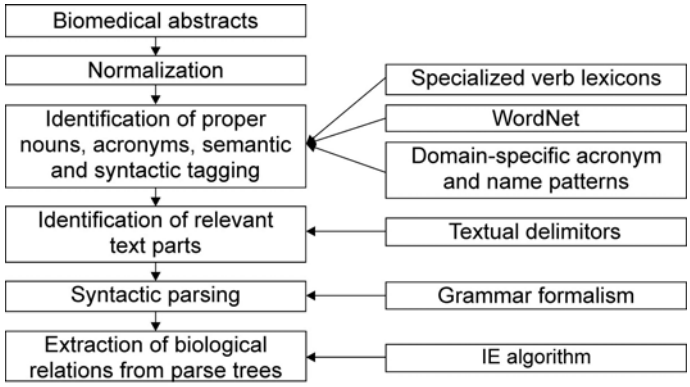


Figure 1. Overview of the information extraction system (Gawronska *et al*., 2005b). Biomedical abstracts are normalized (tokenized) by conversion to plain text, removal of illegal symbols, markup of headings, splitting of the text into sentences, and of sentences into single words. Named entity recognition then identifies proper nouns and acronyms, and semantic and syntactic tagging is performed. A delimitation algorithm then identifies text parts relevant for extraction of relations and a Referent Grammar approach (inspired by Lexical Functional Grammar) is used for syntactic parsing. To the resulting parse trees, an information extraction algorithm is applied in order to identify biological relations representing pathway information. All steps are described in more detail in our previous papers, such as (Gawronska *et al*., 2005b).

MATERIALS AND METHODS

Relevant abstracts were identified by the PubMed search query "crt/dre OR mycr OR dreb1" and the following settings under the "Limits" tab: "Published in the last: 10 years", "Language: English". This gave hits in 58 abstracts which were downloaded and submitted to the normalization module (Fig. 1). The normalized and tagged texts contained approximately 500 sentences and 10 000 words. After identification of the relevant parts of the abstracts, 374 sentences remained and were submitted to the syntactic parser. For successfully parsed sentences, the resulting trees were subjected to information extraction, which attempts to derive a representation of biological relations expressed by each sentence. Only abstracts were used (rather than full texts), since these often contain short and informative summaries of the most relevant and specific findings of the study, while the full text contains large amounts of information that is irrelevant for the purpose of information extraction.

RESULTS AND DISCUSSION

The grammatical knowledge used in the text analysis enables correct interpretation of passive constructions, negations, and certain coordinated constructions where some syntactic elements have been omitted. For example, in a sentence like: "*Expression of the Wcbf2 gene was induced rapidly by low temperature (LT) and drought but not by abscisic acid (ABA)*" the main predicate is represented in the parse tree as:

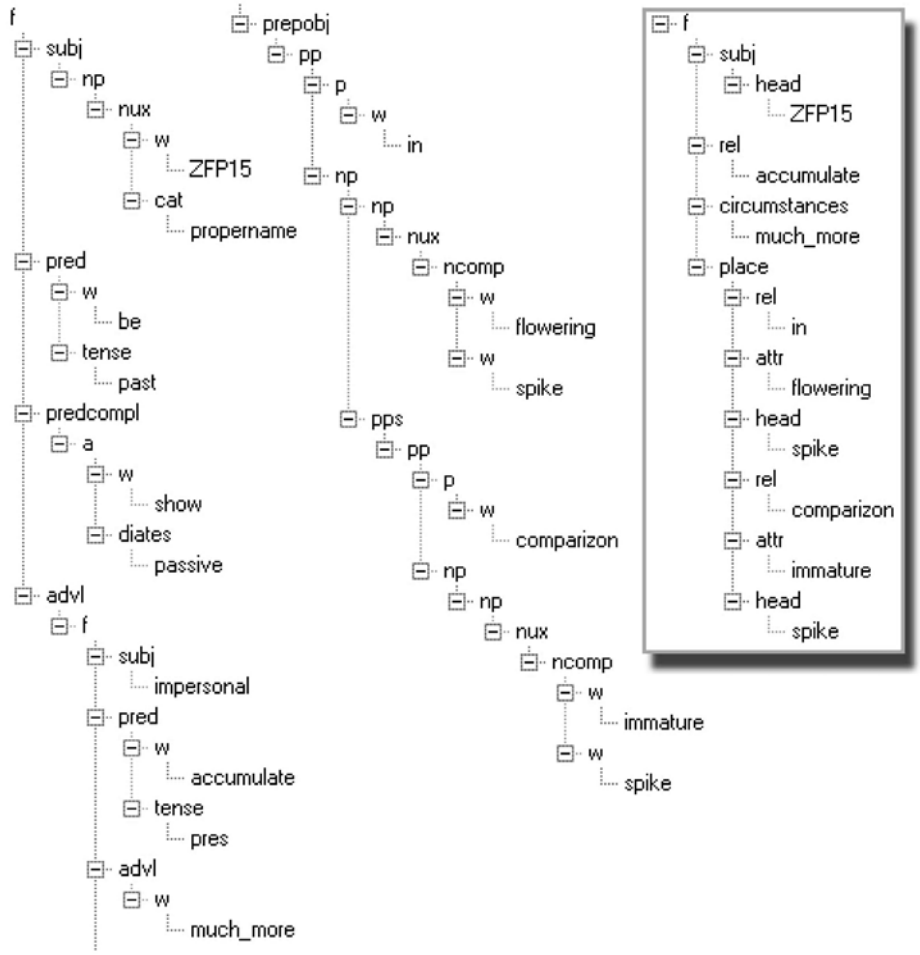


Figure 2. Left: Parse tree generated by syntactic analysis of the sentence *"ZFP15 was shown to accumulate much more in flowering spike than in immature spike"*. Right (in box): Graphical representation of the extracted biological relations. This representation is achieved by deletion of neutral cognition and communication verbs (here: *was shown*), moving the syntactic subject *(ZFP15)* to the co-indexed node ("impersonal") in the infinitive clause, removing some purely grammatical information, and replacing certain syntactic labels by semantic ones, e.g. *pred(irate)* and *p(reposition)* by *rel(ation)*, and *advl (adverbial)* by *circumstances* and *place*.

pred([w(be),nb(sg),tense(past)]),
advl(w(rapidly)),predcompl([a([w(induce),diates(passive)])]).

This enables subsequent use of the information about the predicate's passive voice for appropriate identification of the direction of the induction relation (from *"low temperature"* and drought to *"Wcbf2 expression"*). The omitted subject and predicate of the "but"-clause are co-indexed with the subject and predicate of the first clause (*"Wcbf2"* and *"induction"* in passive voice), and the negation is recognized. Thus, the induction relation between ABA and Wcbf2 is correctly represented as not true. Also quite complicated sentence structures were in most cases parsed correctly, e.g. *"The cultivar difference in freezing tolerance developed during different stages of cold acclimation can be at least partly explained by the differential and coordinated regulation of the predicted Cor/Lea gene signal transduction pathway that is mediated by the CBF/DREB1 transcription factors in common wheat."* Here, the main predicate was correctly identified as a modal verb followed by an infinite clause: pred([w(can)]),infobj(f(subj(impersonal),pred([w(be),[], tense(inf)]),advl (w(at_least)), predcompl([adv(w(partly)), a([w(explain),diates(passive)])]).

The embedded predicates are correctly connected to their arguments: *"The predicted Cor/Lea gene signal transduction" pathway is interpreted as being exposed to "the differential and coordinated regulation", and as being mediated by "the CBF/DREB1 transcription factors".*

Overall, 63 % of the sentences were successfully parsed. This rate is somewhat lower than in previous evaluations (Gawronska *et al*., 2005b), which seems to be due to the presence of lexical gaps and the lack of patterns for Latin plant names in the current lexicon. The parser has access to the most common patterns for gene and protein names, but it does currently not recognize certain Latin abbreviations, like *Capsicum annuum*. We also observed a few cases of incorrect sentence delimitation during the normalization procedure, which is related to the previously mentioned problem. Given a phrase like *Capsicum annuum* L. *cv*. *Pukang*, the normalization procedure interprets "cv" as the last word of a sentence, rather than an abbreviation. Both of these problems can clearly be solved simply by extending the lexicon.

In comparison with previous evaluations of the system (Gawronska *et al*., 2005b) we here found more examples of very long ambiguous coordinated structures. In particular, the parser had difficulty with those involving participle clauses and prepositional phrases attached to the coordinated constituents, e.g. in the sentence: In parallel with these changes, ***increases in photosynthetic efficiency and capacity, pigment pool sizes, increased capacities of the Calvin cycle enzymes, and enzymes of starch and sucrose biosynthesis, as well as glycolysis and oxaloacetate/malate exchange*** *are seen, suggesting that BNCBF overexpression has partially mimicked cold-induced photosynthetic acclimation constitutively.* However, such complex sentences are difficult to interpret even for human readers.

REFERENCES

Bräutigam M. *et al*. (2005) Generation and analysis of 9792 EST sequences from cold acclimated oat, Avena sativa. *BMC Plant Biol*., **5**, 18.

Bräutigam M. *et al*. (2006) Development of a Swedish winter oat with gene technology and molecular breeding. *J. of the Swedish Seed Association*, (1-2).

Gawronska B. *et al*. (2004) Natural language technology in multi-source information fusion. *Proc. of Intl. IPSI-2004k Conference*, Kopaonik, Serbia.

Gawronska B. *et al*. (2005a) Syntactic, semantic and referential patterns in biomedical texts: towards in-depth text comprehension for the purpose of bioinformatics. *Proc. of 2nd Intl. Workshop on Natural Language Understanding and Cognitive Science*, Miami, FL, USA.

Gawronska B. *et al*. (2005b) Tracking biological relations in text: a referent grammar-based approach. *Proc. of Biomedical Ontologies and Text Processing*, Madrid, Spain.

Hirschman L. *et al*. (2002) Accomplishments and challenges in litterature data mining for biology. *Bioinformatics*, **18**(12), 1553–1561.

Kanehisa M. *et al*. (2004) The KEGG resources for deciphering the genome. *Nucl. Acids Res*., **32**, D277–D280.

Olsson B. *et al*. Deriving pathway maps from text using a grammar-based approach. *J. of Bioinf. and Comp. Biol*. (in press).

Thomashow m.f. (1999) Plant cold acclimation: freezing tolerance genes and regulatory mechanisms. *Ann. Rev. in Plant Phys. and Plant Mol. Bio*., **50**, 571–599.

THE ONTOLOGY OF ECOSYSTEMS

***Suslov V.V.*[*1], *Sergeev M.G.*[2], *Yurlova N.I.*[3], *Miginsky D.S.*[1, 2]**
[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia
[*] Corresponding author: e-mail: valya@bionet.nsc.ru

Key words: ecosystems, gene networks, database presentation formats, ontologies

SUMMARY

Motivation: The ontology of a knowledge domain is a set of statements that are true, no matter what circumstances may arise in that knowledge domain. The aim of this work is to develop a formal ontological description of the structural and functional organization of ecosystems (which all together comprise the knowledge domain) with due regard to their possible dynamics/development/evolution. Since ecology is an essentially synthetic science (Begon *et al*., 1989; Margalef, 1992), multidisciplinary research is a necessity. However, what happens is that individuals working in different fields stick to their respective paradigms and terminology, which brings up a terminology mess and hampers comparison (Sergeev *et al*., 2006). Providing a uniform format for description and comparison, this formalization was originally developed in response to a need of developing a strategy for the creation of a database for an all-round description of biosystems (Miginsky *et al*., 2006). Now that a major portion of this task has been completed, the ontology can be used for navigation in this database.

Results: The definition of an econetwork is provided; the ontology of their micro- and macrodynamics is described. The ontology provides eco- and gene networks linkage.

INTRODUCTION

There is a wealth of empirical data on the organization of ecosystems and a lot of attempts at making generalizations about the available material (Begon *et al*., 1989; Allen, Hoekstra, 1992). The main factors which keep scientists trying are as follows: a true complexity of organism communities, inert and biologically inert components (Begon *et al*., 1989); deficient integrity of the ecosystems due to the lack of a data storage like the genome (Dajoz, 1976; Margalef, 1992), which is why the community structure and characteristics are subject to variation; a conflict of different ecosystem description formats, due to which the knowledge domain is fuzzily defined and the ontology may not be developed by definition. There are three approaches in use to describe ecosystems: 1) *functional*, which was pioneered by Tansley (1935) and later formalized in the form of trophic networks (Luczkovich, 2003); 2) *geoecological*, extensively formalized in the works of Sochava (1978). Here a biogeocenosis is treated as an elementary dynamic coevolving system of live, inert and biologically inert elements. Unlike trophic networks, this system has a certain dimensionality (Sukachev, 1972); 3) *paragenetic*, based on correlations of the creatures ontogeny programs in a community (Razumovsky, 1981). This was first developed to be used in parasitology (Pavlovsky, 1934; Beklemishev, 1970a). Common to these three approaches is that they share the concept of *an elementary object* (*EO*) – which is a minimal fundamental object which cannot be broken down into anything smaller without loss of properties. EOs can form groups of a certain

dimensionality, or scale, which establish various hierarchical relationships depending on the physical properties of the EOs and the approach used (Zherikhin, 1994).

There are *elementary interactions* (EIs) between the EOs and the groups they are in. The EIs provide the exchange of matter, energy and information. The EI classification is poorly developed compared to that of EO. The most detailed classification by Beklemishev (1970b) is poorly formalized. The author argues that the same EI can be described by different classes or switched to a different class when the dimensionality/size changes. Given the settings, the development of the ontology is equivalent to EO formalization, no matter what descriptive approach or EI classification quality.

METHODS AND ALGORITHMS

The ontology was tested for applicability on two different ecosystems: parasitic, exemplified by the lifecycle of the trematode *Echinoparyphium aconiatum*, and steppen, by nongregarious grasshoppers (Sergeev *et al*., 2006). Data on the structure of network interactions in the course of the complex life cycle of the trematode *Echinoparyphium aconiatum* had previously been made available (Vodyanitskii *et al*., 2002) to the GeneNet database (Ananko *et al*., 2002). Data on the ecosystems comprising grasshoppers were collected by annotation of scientific publications (guided by Sergeev).

RESULTS AND DISCUSSION

An ecosystem can formally be rendered as a graph with EOs (no matter what properties) or EO groups as the nodes connected by EIs as the edges. Call this representation an *econetwork*. Two sorts of change can occur in econetworks: 1) *microdynamics*, changes in the strength of edges and nodes due to change in the values of variables characterizing EOs and EIs; 2) *macrodynamics*, qualitative changes to the graph due to the emergence or extinction of EOs and EIs. An example of microdynamics is population dynamics; a striking example of macrodynamics is seasonal communities rearrangements[3].

As a microdynamic event, the ontology (Fig. 1*a*) distinguishes between microdynamic EOs (m1) and microdynamic EIs (m2). We classify EOs into the following subentries: live (m1.1), organisms and their factions: populations, guilds, synusia, and so on; inert (m1.2), landscape elements, chemicals, physical factors); biologically inert (m1.3), soils, silts, dead matter and biologically active compounds released by organisms to the environment (pheromones, vitamins, metabolites). Microdynamic EIs[4] (m2) can affect an EO by way of affecting its characteristics or location (m2.2, reaction) or can influence another EI (m2.1, regulatory effect[5]). We classify microdynamic EIs into mass/energy-

[3] Rank dynamics, which is the common lowermost level, should be mentioned (see Sergeev *et al*., 2006). The rank is a numerical characteristic of an EO such that its change may not affect the EO, but what it can do is strongly affect the interaction between other EOs in the eco- or gene network. For example, the parasite burden on an infected mollusk can be described by the following ranks: Moll[s,re,m], where s is the sporocysts number, re is the rediae number, m is the metacercariae number. In ecosystems, ranks can be used to describe, for example, age-specific change in body size, biocenosis area change; in gene networks, to describe modification of the protein phosphorylation degree, DNA methylation degree, etc.

[4] Micro- and macrodynamic EIs can be immediate (without intermediate stages) or mediate. For simplicity, this level of division is not present in Fig. 1*a*, *b*.

[5] The agreed classification for the micro- and macrodynamic regulatory effects is in part identical to that used in GeneNet (Ananko *et al*., 2002): increase, decrease, switch-on, switch-off; but there is a difference too. Microdynamic regulatory effects are divided into common ones (M.2.1.1), which are

related (m2.2.1), topic (m2.2.2), fabric (m2.2.3), phoric (m2.2.4)[6] and developmental (m2.2.5). If the effect is directly proportional to the amount of substance (energy) consumed, the EI is trophic (m2.2.1.1); if the effect is strongly non-linear, the EI is information-related (m2.2.1.2) (Fig. 1*a*), among which we place tactile, visual, immune and the like interactions. Topic EIs (m2.2.2) are limited to competition for room within habitats (for example, competition for room by helminthes in the limited space of the host's body cavities), conditioning of biotope characteristics by various elements of the biogeocenosis[7] and physical limitations on the accessibility of the elements of biogeocenosis. Phoric EIs (m2.2.4) are EO relocations from structure to structure without much change (m2.2.4.1): active or passive relocations of live EOs[8]; transport EIs (m2.2.4.2): essentially the same as phoric but relate to inert EOs. Fabric EIs (m2.2.3): the activity of organisms leading to a physical rearrangement of the biotope or biocenosis (for example, hole digging by soil and silt dwellers[9]). Developmental (m2.2.5), which are the rearrangements that occur to EOs for intrinsic reasons: ontogenetic EIs for live EOs (m2.2.5.1); genesis-related EIs for inert and biologically inert EOs[10] (m2.2.5.2). Biochemical ontogenetic EIs (m2.2.5.1.1) provide a link between ecosystems and gene networks: they open access to genes related to ontogenesis. Linkage with gene networks can be envisaged for more EIs of live EOs as more data accrue.

Macrodynamic EOs are the stages (M1) in the history of ecosystems (Fig. 1*b*). At each stage, the ecosystem parameters display stability (the same number of species, the biotope elements are all there, the same temperature, humidity and so on). It is this stability that accounts for there being stages, hence a stable graph of connections.

We classify the stages into exogenous (M1.1), which can be primary (M1.1.1) or secondary (M1.1.2)[11], and endogenous (M1.2). Why is the history of ecosystems staged? The underlying mechanisms of being staged are the property of the ecosystem itself (a misbalance of biogeochemical cycles, the gene networks of edifying species). A change to the econetwok graph due to the addition or elimination of a community member describes the cenotic stage (M1.2.1); A change to the econetwok graph without change in the econetwok community describes the clinal stage (M1.2.2). Changes to the econetwok graph due to the ontogenesis of live microdynamic EOs or the genesis of inert microdynamic EOs describe developmental stages (M1.2.3): ontogenetic (M1.2.3.1) and genesis-related (M1.2.3.2), respectively. Macrodynamic EIs (M2) include regulatory effects (M2.1) and reactions (M2.2), which are essentially the transitions between stages and can be either exogenous (M2.2.1)[12] or endogenous (M2.2.2). The endogenous ones include cenotic (M2.2.2.2.1), clinal (M2.2.2.2.2) and developmental transitions (M2.2.2.3) which are ontogenetic (M2.2.3.1) and genesis-related (M2.2.3.2) (Fig. 1*b*).

---

classified as per GeneNet, and those possessing an optimum (M2.1.2), which depend on the "strength", or the "amount", of the affecting object in the econetwork. The optimum number of parasites enhances the homeostasis of the ecosystem, while any number up or down disrupts it (Fig. 1*a*, *b*).

[6] These terms were borrowed from Beklemishev (1970b), but our interpretation differs from the original.

[7] The example of conditioning is a shading. Poikilothermic, grasshoppers often travel between open, warmed microstations and microstations covered with vegetation where they feed (Sergeev *et al*., 2006).

[8] The encysted metacercariae *E. aconiatum* present in mollusks are liable to a peculiar sort of passive migration: being swallowed by birds who feed on mollusks (Sergeev *et al*., 2006).

[9] If this activity brings up a new biotope, the ontology of macrodynamic events should be applied to.

[10] Examples of ontogenetic EIs are growth, ontogeny, lifecycle, behavioural instinct. Examples of genesis-related EIs are soil genesis, sedimentogenesis, dead matter decomposition.

[11] The primary stages are classified based on the astronomic events they are associated with (Dajoz, 1976). For simplicity, only circadian (M1.1.1.1) and seasonal stages (M1.1.1.2) are shown, while the others are in the astronomic group (M1.1.1.3) (Fig. 1*b*). The secondary stages are of extremely heterogenic origin and so their classification is beyond the scope of the present ontology.

[12] For simplicity, the division into primary and secondary EIs is not shown (Fig. 1*b*).

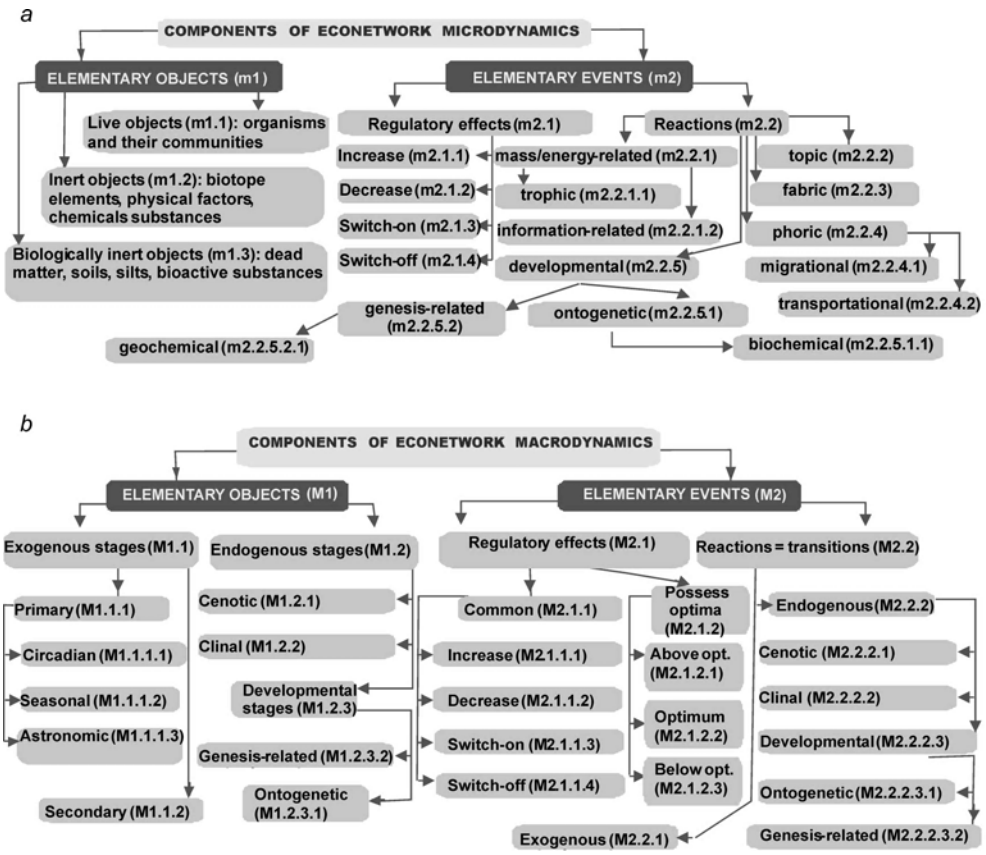


Figure 1. Components of the microdynamics (*a*) and macrodynamics (*b*) of econetworks.

ACKNOWLEDGEMENTS

The authors are grateful to N.A. Kolchanov for a fruitful discussion.

The work is supported by RFBR grants Nos 03-04-48506-a, 03-01-00328; integration SB RAS projects Nos 119, 142, 145, 148; RAS projects of Biodiversity (12.4), Physics and Biology (10.4); project from RAS Presidium"Origin and Evolution of the Biosphere".

REFERENCES

Allen T.F.H., Hoekstra T.W. *Toward a unified ecology.* New York *a.o.*: Columbia Univ. Press, 1992. 383 p.

Ananko E.A., Podkolodny N.L., Stepanenko I.L. *et al*. (2002) GeneNet: a database on structure and functional organisation of gene networks. *Nucl. Acids Res.*, **30,** 398–401.

Beklemishev N.V. (1970a) *Parasite and Nidicole Populations and Micropopulations. Biotechnological fundamentals of Comparative Parasitology.* Moscow: Nauka, 215–226. (in Russ.).

Beklemishev N.V. (1970b) *On Classification of Biocenotic Interactions,* Ibid. 90–138 (in Russ.).

Begon M., Harper J., Townsend C. (1989) *Ecology.* In 2 volumes Moscow: Mir. (in Russ).

Dajoz R. (1976) *Précis d'écologie.* Moscow: Progress, 450 p. (in Russ.).

Luczkovich J.J., Borgatti S.P., Johnson J.C. *et al*. (2003) Defining and measuring trophic role similarity in food webs using regular equivalence. *J. Theor. Biol.*, **220(**3), 303–321.

Margalef R. (1992) *The Looks of the Biosphere.* Moscow: Nauka, 214 p. (in Russ.).

Miginsky D.S., Suslov V.V., Rasskazov D.A. *et al*. (2006) Architecture of software toolkit for storing and operatng with biosystems models. *This issue.*

Pavlovsky Eu.N. (1934) The Organism as a habitat. *Priroda*, **1**, 80–91. (in Russ.).

Razumovsky S.M. (1981) *On the Dynamics of Biogeocenoses.* Moscow: Nauka, 231 p. (in Russ.).

Sergeev M.G., Suslov V.V., Miginsky D.S. *et al*. (2006) Towards a database for ecosystem description using network technology. *Ecosystem biodiversity and dynamics: informatics-based approaches and modeling*. Novosibirsk: SB RAS Publisher, in press. (in Russ.).

Sochava V.B. (1978) *Introduction to Geosystems*. Novosibirsk: Nauka, 319 p. (in Russ.).

Sukachev V.N. (1972) *Fundamentals of Wood Typology.* Leningrad: Nauka. 418 p. (in Russ.).

Tansley A. (1935) The use and abuse of vegetational concepts and terms. *Ecology*, **16,** 284–307.

Vodyanitskii S.N., Yurlova N.I., Suslov V.V. Formal description of the trematode ecoparasitic system on using the GeneNet data format. Proc. of the 2nd Conference on Bioinformatics of Genome Regulation and Structure, BGRS, 2002, Novosibirsk, ICG. 2002, **2,** 210–213.

Zherikhin V.V. (1994) Evolutionary Biocenology. The Choice of Models. *Ecosystem Rearrangements and Biosphere Evolution.* Moscow: Nedra, 13–20. (in Russ.).

THE GArna TOOLBOX FOR RNA STRUCTURE ANALYSIS: THE 2006 STATE OF THE ART

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

Key words: RNA secondary structure, inverse folding, genetic algorithm, miRNA, mRNA, translation

SUMMARY

Motivation: RNA-RNA interactions control numerous processes in cell. There is a continuing demand for computer programs based on analysis of RNA to study RNA structure-function relationships, develop new drugs, analyze genomic sequences.

Results: Here we describe a program package for analysis of RNA-RNA interactions. Most package options are available through the Internet (Titov *et al*., 2006): simulation of RNA folding, search of RNA fitted to a given secondary structure, calculations of the secondary structure using additional information. GArna provides tools for drawing RNA secondary structure in a fairly quick and easy way. The program package was recently extended with new programs to address the following issues: calculation of RNA secondary structure using energy minimization, design of mRNA 5′UTRs, miRNA gene design, calculation of RNA helix dynamics, design of siRNAs.

Availability: wwwmgs2.bionet.nsc.ru/mgs/systems/garna or on request.

INTRODUCTION

RNA structure defines its function and interactions with other molecules. Algorithms are available via the Internet that makes feasible the calculation of the RNA secondary structure (Hofacker, 2003; Zuker, 2003). Although the thermodynamical algorithms (Hofacker, 2003; Zuker, 2003) remain most wide spread, they can be outperformed by evolutionary programming approach (Wiese, Hendriks, 2006). We have previously reported the GArna server to solve this and inverse problems (Titov *et al*., 2006), based on genetic algorithm. Modern biotechnology has made practicable the construction of new molecules to expand the functional repertoire of RNA and optimize the processes with its involvement. A recent hot topic of ribonomics is small RNAs. They revealed new horizons for RNA control of cell processes. More and more is known about the role of RNA and there is a growing need in expanding the existing packages by new programs. In this paper we briefly describe the programs for RNA analysis implemented in the GArna package. The programs are divided into three groups:

1) Sequence analysis: search of potentially structured regions in long RNAs; search for miRNA sites that repress translation.

2) Calculation of RNA secondary structure: search of the optimal and suboptimal secondary structures; folding simulation; calculation of oligonucleotide-RNA structure; calculation of the RNA secondary structure using experimental data; calculation of the dynamical properties of RNA helixes.

3) RNA design: RNA design with a given secondary structure; design of the optimal mRNA 5′ UTR; miRNA gene design; design of siRNAs.

The programs for search of potentially structured regions in long RNAs, RNA folding simulation, calculation of oligonucleotide-RNA complex structure, calculation of the RNA secondary structure using experimental data, RNA design with a given secondary structure have been described elsewhere (Titov *et al*., 2006). The others, performing the search of miRNA sites that repress translation, calculation of the dynamical properties of RNA helixes, design of the optimal mRNA 5′ UTRs, of miRNA genes and of siRNAs to suppress translation have been developed since last report (Titov *et al*., 2006).

SEQUENCE ANALYSIS

Search of potentially structured regions in long RNAs.

It has described elsewhere how the program performs this function (Titov *et al*., 2006). The program calculates the nucleotide score that strongly correlates with the secondary structure energy. The matrix for the potential structure and the score profile are displayed at the output. The program predictions can be used for further experimental (Napierala *et al*., 2005) or computational study of found regions.

Search of miRNA sites that repress translation.

Search is based on the model of the translational arrest by the RNP complex. (for details, see (Titov, Ivanisenko, 2006). The miRNA and mRNA sequences are input, and the program outputs the potential repression sites, if detected.

CALCULATION OF RNA SECONDARY STRUCTURE

Search of the optimal and suboptimal secondary structures.

The program is based on the modified recursive algorithm (Zuker, 2003) and it calculates optimal and suboptimal set of structure in a given energy window.

RNA folding simulation

The program relies on a rapid genetic algorithm and it allows to find the RNA folding intermediates (for details, see (Titov *et al*., 2002)). The program outputs the energy, the Z-score of the RNA sequence and the RNA structure graph. Drawn RNA secondary structure may be moved and scaled with the mouse.

Calculation of oligonucleotide-RNA complex structure.

This calculation is offered as an option for the preceding program. It is the user's choice to set the region where RNA interacts with the oligonucleotide. The program can calculate the structure provided that the target region is screened from intramolecule interactions.

Calculation of the dynamical properties of RNA helixes.

The user sets the helix sequence. The program calculates the partition function of the helix and reports the base-pairing probabilities, one position after another.

Calculation of the RNA secondary structure using experimental data.

By using the point-and-click interface, the user sets the paired or unpaired bases. The program minimizes the structure energy with the given constraints ((Titov *et al*., 2006). After calculation, the user can return to the editor window, modify constraints, and reiterate calculations (Fig. 1). The program may be useful in completing definition of a partly known structure.

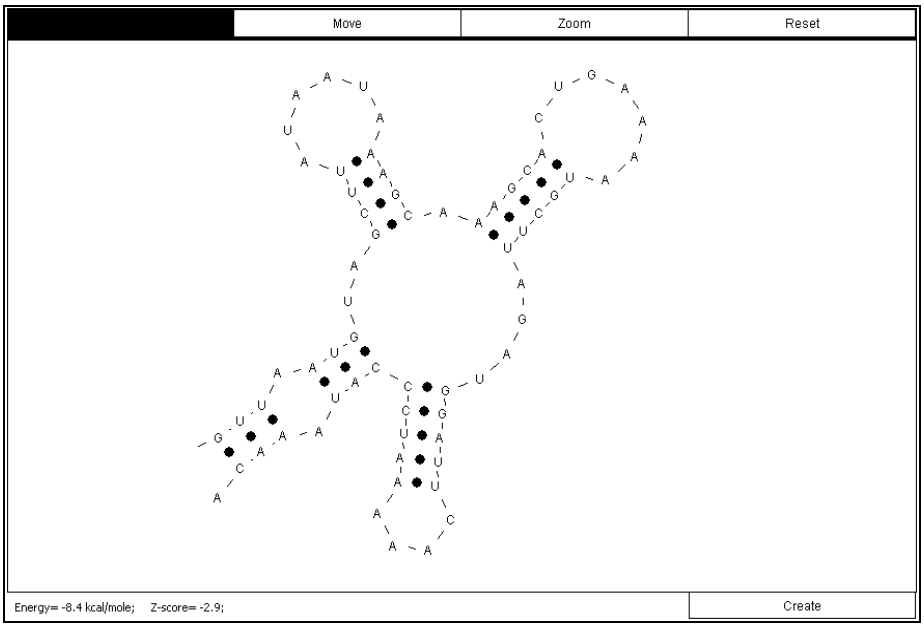


Figure 1. An example of constrained optimization by our program: 67nt sequence was forced to fold into cloverleaf structure.

RNA DESIGN

RNA design with a given secondary structure.

The program interface is similar to the preceding. The user sets the secondary structure. The program searches the RNA sequence that maximizes the thermodynamic probability of the target structure.

Design of the optimal mRNA 5' UTR.

As known, various properties of mRNA leader region can affect translational efficiency. These include the competing AUG codons, the nucleotide and secondary structure environment of the start codon and other properties. The user inputs the mRNA leader sequence. The program can stepwise optimize it by nucleotide substitutions as the user wishes. The program can remove false translation starts, optimize the nearest context of AUG codon, get rid of the leader secondary structure, create the stable hairpin by synonymous substitutions for ribosome pausing. The user may set the nucleotide positions that must remain unaltered.

miRNA GENE DESIGN

The miRNA gene is stable hairpin that contains miRNA sequence. The energy of the secondary structure and the probability to reside in hairpin form are the two guidline features in the search of miRNA genes. The candidate gene and miRNA are user's choice. The program restores miRNA complementarity within the gene and optimizes the thermodynamic characteristics of the gene hairpin.

siRNA design to suppress translation.

The procedure is based on the model of translational arrest by the RNP-mRNA complex. (Titov, Ivanisenko, 2006). The model assumes that a complex of RNP-particles, with each attached to the siRNA-mRNA binding site, supresses the translation. The specificity of complex formation depends on the interaction of siRNA with the targets and

mRNA structure in-between. The program searches for siRNA, that maximizes the probability of RNP complex formation, and outputs the siRNA candidate list.

CONCLUSION

Our package is developing as a continuous effort of RNA group from Laboratory of Theoretical Genetics of Novosibirsk Institute of Cytology and Genetics. We extended it with a number of new programs since last report; these programs will be installed on web-server GArna (wwwmgs2.bionet.nsc.ru/mgs/systems/garna).

ACKNOWLEDGEMENTS

The work was supported by Innovation Project of Federal Agency of Science and Innovation 02.434.11.3004 “Identification of perspective targets for new medicine drugs based on gene network reconstruction”.

REFERENCES

Hofacker I.L. (2003) Vienna RNA secondary structure server. *NAR*, **31**(13), 3429–3431.

Napierala M., Michalowski D., de Mezer M., Krzyzosiak W.J. (2005) Facile FMR1 mRNA structure regulation by interruptions in CGG repeats. *NAR*, **33**(2), 451–463.

Titov I.I., Voroviev D.G., Ivanisenko V.A., Kolchanov N.A. (2002). Fast genetic algorithm for RNA secondary structure analysis. *Russ. Chem. Bull*., 7, 1047–1056.

Titov I., Vorobiev D., Palyanov A. (2006) A toolbox for analysis of RNA secondary structure based on genetic algorithm. In Kolchanov N., Hofestaedt R. (eds), *Bioinformatics of Genome Regulation and Structure II.* Springer Science+Business Media, Inc., pp. 105–110.

Titov I.I., Ivanisenko A.Yu. (2006) A model of the translational inhibition by miRISC complex describes protein synthesis variation induced by mutations in mammalian miRNA sites. *This issue*.

Wiese K.C., Hendriks A. (2006) Comparison of P-RnaPredict and mfold—algorithms for RNA secondary structure prediction. *Bioinformatics*, **22**(8), 934–942.

Zuker M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *NAR*, **31**(13), 3406.

A MODEL OF THE TRANSLATIONAL INHIBITION BY MIRISC COMPLEX DESCRIBES PROTEIN SYNTHESIS VARIATIONS INDUCED BY MUTATIONS IN MAMMALIAN miRNA SITES

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

Key words: RNA secondary structure, miRNA, siRNA, RNP, RISC, translation, hepatitis C virus

SUMMARY

Motivation: Most miRNA sites possess imperfect complementarity and presumably play a role in mRNA translational inhibition without degradation. The mechanism of translational inhibition of miRNA is poorly studied, because this study requires proteome profiling. An insight into how multiple sites function is important in choosing a scenario for siRNA design and reducing off-target effects.

Results: We propose a model for translational inhibition by miRISC complex (miRNA-containing RNA induced silencing complex (Kim, Nam, 2006)). In this model, miRNP particles cooperatively bind to an mRNA template, and miRNA mediate the interactions. A score based on rough calculations of the energy of the complex is proposed as a measure of the efficiency of translational inhibition. The calculations suggest that most of what are thought as potential targets, in fact, are not due to inappropriate relative position of miRNA sites. A comparison with experimental data (Lewis *et al*., 2003) demonstrates that the score describes protein synthesis variation induced by mutations in mammalian miRNA sites. We have performed a genome-wide scan of the hepatitis C virus for human miRNA sites and identified what siRNA appear to inhibit virus translation.

Availability: available on request.

INTRODUCTION

Presumably, miRNA exerts regulation on as much as 10 % of human genes (John *et al*., 2004). The mechanism of gene expression regulation for the case that miRNA nearly perfectly bind to complementary sites and promote mRNA degradation is well studied. When potential sites possess a lesser complementarity, the activity of miRNA as regulators of gene expression is believed to be related to translational inhibition. Only occasional studies of the molecular mechanism of translation inhibition have been performed to date: they are difficult to carry out, because proteome profiling is required. Presumably, miRNA mediate assembly of the miRNP complexes on mRNA, which represses translation. There are two points that need to be clarified. First, since short complementary regions for miRNA, not serving as the real miRNA sites, can be found in any mRNA, what stability of miRNA-mRNA duplexes is critical? Secondly, what should mRNA structure and relative position of the sites be like to provide the successful

assembly of an RNP complex? To find out, we have developed a physical-chemical model of the complex. We have demonstrated that the results of predictions using the proposed model are consistent with data from mutational experiments (Lewis *et al*., 2003). Based on the model, we have developed software programs as tools for siRNA activity prediction and siRNA design. We have performed a genome-wide scan of the hepatitis C virus for human miRNA sites and identified siRNA that appear to be inhibitor of virus translation. These software programs form part of a program package for analysis of RNA secondary structure (Titov, 2006).

RESULTS

To help discriminate between a tremendously large number of sites with little complementarity from those that really work, we proposed a model for miRNP complex assembly (Fig. 1).

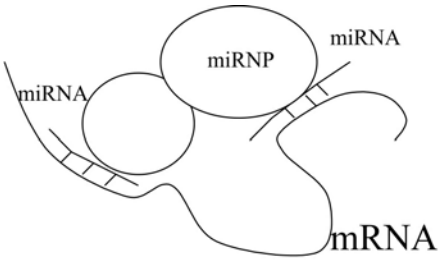


Figure 1. A model for miRNP complex assembly on mRNA. miRNAs guide mRNA-protein assembly indicating where to bind to the template. As protein-protein interactions dominate, the assembly is irreversible.

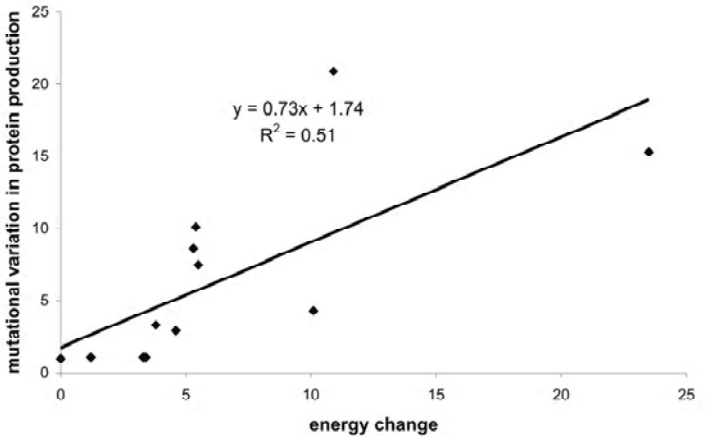


Figure 2. The variation of the observed amount of the protein due to mutations in mammalian miRNA binding sites of mRNA (Lewis *et al*., 2003), normalized to the WT amount, correlates with the change in the calculated stability of the miRNP complex (Eq.1), $r^2 = 0.51$.

We proposed that the last stage of assembly, with protein-protein interactions under way, is irreversible. If this is true, then the assembly of the miRNP complex should depend on the stability of interactions between nucleic acids: (a) intermolecular interactions between miRNA and mRNA and (b) intramolecular interactions within mRNA. We assumed that other interactions were either weak or constant and omitted

them from consideration. The total energy of interactions (a) and (b) was calculated using Turner's thermodynamic table (Jaeger *et al*., 1989) and summing over all interactions:

$$\Delta G = \sum_a \Delta G_a + \sum_b \Delta G_b. \quad (1)$$

Model verification was performed using experimental data on multiple mammalian miRNA sites. In the experiment miRNA-mRNA interactions were disrupted resulting in variation in protein production (Lewis *et al*., 2003). Our approach describes well the observed change in protein synthesis (Fig. 2). For most wild type sites (Lewis *et al*., 2003), the total interaction energy calculated by Eq.(1) was between –30 and –40 kcal/mole (data not shown). This suggests that the energy value at –30 kcal/mole of nucleic-nucleic interactions is critical for effective translational inhibition by miRNP complex.

DISCUSSION

A model decribed above can help design of siRNA for translational inhibition of the hepatitis C virus. A genome-wide scan of the hepatitis C virus as a target for human miRNAs using the software program (http://vita.mbc.nctu.edu.tw) and genome sequence H77 (http://hcv.lanl.gov/components/hcv-db) revealed that only weak human miRNA sites were present, which might be due to the evolutionary adaptation of the virus to the host organism. We have developed a software program for the design of efficient siRNAs; this tool forms part of a program package for analysis of RNA secondary structure (Titov, 2006). The program identifies which siRNA sequence is the best match to a particular user-defined mRNA sequence. The program generates a list of siRNA ranked in order of descending score (in order of decreasing probability that miRNP particles aggregate into a complex according to our model). Only one pair of siRNA sites has been identified as having the energy below the threshold (–30 kcal/mole). The sites are confined to the 3′ region of the virus genome, and the calculated energy is considerable, –52.5 kcal/mole. Thus, the proposed siRNA may appear to be inhibitors of virus translation.

ACKNOWLEDGEMENT

This work was supported by Innovation project of Federal Agency of Science and Innovation 02.434.11.3004 "Identification of perspective targets for new medicine drugs based on gene network reconstruction".

REFERENCES

John B. *et al*. (2004) Human microRNA targets. *PLoS Biol*., **2**(11), e363.
Kim V.N., Nam J.W. (2006) Genomics of microRNA. *Trends Genet*., **22**(3), 165–173
Lewis B.P. *et al*. (2003) Prediction of mammalian miRNA targets. *Cell*, **115**, 787–798.
Titov I.I. (2006) The GArna toolbox for RNA structure analysis: the 2006 state of the art. *This issue*.
Jaeger A.H., Turner D.H., Zuker M. (1989) Improved predictions of secondary structures for RNA. *Proc. Natl. Acad. Sci. USA*, **86**, 7706–7710.

PART 7. SHORT ABSTRACTS

DETERMINATION OF NATIVELY UNFOLDED REGIONS OF SUPEROXIDE DISMUTASE FROM *PACIFASTACUS LENIUSCULUS*

***Cavas L.*[*1], *Cavas C.K.*[2]**
[1] Dokuz Eylul University, Faculty of Arts and Sciences, Department of Chemistry, Division of Biochemistry; [2] Dokuz Eylul University, Faculty of Arts and Sciences, Department of Statistics, 35160, Kaynaklar Campus, IZMIR/TURKEY
[*] Corresponding author: lcavas@hotmail.com, levent_cavas@yahoo.com

The roles of natively unfolded proteins in some important biological functions such as transcriptional regulation, translation and cellular signal transduction have been shown in many papers. According to the results of these papers on unfolded regions of proteins, having unfolded regions provides proteins better flexibility and also better interaction ability with their ligand or binding sites. Superoxide dismutase (SOD) is a house-keeping enzyme that protects cells from harmful effects of superoxide radical anion. SOD catalyses the dismutation reaction of superoxide radical anion to hydrogen peroxide. In the present study, unfolded regions of a novel cell surface SOD that has been found in *Pacifastacus leniusculus* were analysed by a newly developed bioinformatic tool called RONN. Amino acid composition and some physicochemical parameters of SOD from *Pacifastacus leniusculus* such as pI, negatively and positively charged residues, aliphatic index and grand average of hydropathicity were also investigated in the present study. Although there are technically three unfolded regions (at residues 88–92, 153–174 and 183–217) in SOD from *Pacifastacus leniusculus*, probability of disorder of the last region was higher than those of other unfolded regions. One of the possible property of unfolded regions in proteins is to provide better interaction with their binding sites. Inasmuch as cell surface SOD from *Pacifastacus leniusculus* is a binding protein for a cell adhesive peroxidase in crayfish, having unfolded regions provides SOD to bind effectively cell adhesive peroxidase. Algorithms developed in bioinformatics for prediction for unfolded region in proteins might provide a different view of point for evaluation of structure-activity relationships.

DYNAMIC PROGRAMMING ALGORITHM PARALLIZATION FOR PROTEIN FOLDING

Dulko V.**, *Feranchuk S.
United Institute of Informatics Problems of the National Academy of Sciences of Belarus, Minsk, Belarus
*Corresponding author: e-mail: dulko@inbox.ru

Dynamic programming algorithms are widely used in bioinformatics. In some situations they are very time-consuming and it is necessary to use parallel computations. We consider a problem of protein folding under constrains on the combinatorial search applied. Parallelization of these algorithms is not obvious. We presented an approach to this problem. The first and obvious step is to separate accumulating process from computation processes. An accumulating process should collect all results from computation processes and find a series of best results to be calculated next. But the problem is that the load of the only accumulating process can be high. To solve this problem, a notion of a *front* is introduced as a set of best results for some level of progress of the whole task. This front can be separated for several regions, and one accumulating process is assigned to each region. Computation processes are assigned to each accumulating process and take data from it, while accumulating processes can take results from each computation process. This solution is implemented on supercomputer SKIF K-1000 using MPI library and shows a good performance.

The work was partially supported by the INTAS project No. 04-77-7178.

REFERENCES

Godzik A. (2003) Fold recognition methods, In Bourne P.E., Weissig H. (eds), *Structural bioinformatics,* 525–546 (Wiley-Liss).

Guerra C., Istrail S. (eds) (2003) Protein structure comparison: algorithms and applications, In Guerra C., Istrail S. (eds), *Protein Structure Analysis and Design*, pp. 1–33.

Ferrari C., Guerra C. (2003) Geometric methods for Protein structure comparison. In Guerra C., Istrail S. (eds), *Protein Structure Analysis and Design*, pp. 57–82.

COMPUTATIONAL IDENTIFICATION OF TRANSCRIPTION FACTOR BINDING SITES WITH VARIABLE ORDER BAYESIAN NETWORKS

Grosse Ivo
Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) and Bioinformatics Center Gatersleben-Halle, Gatersleben, Germany
e-mail: grosse@ipk-gatersleben.de

Deciphering genomic sequences is a challenge of the 21st century, and one of our activities focuses on the computational identification of transcription factor binding sites. While tremendous progress has been made along these lines, existing algorithms are often prone to over-fitting, and their accuracies are not yet satisfactory. In an attempt to circumvent the over-fitting problem, we combine Variable Order Markov models with Bayesian networks, and we demonstrate how the resulting Variable Order Bayesian networks can be applied to the identification of prokaryotic and eukaryotic transcription factor binding sites.

HAPLOTYPE BLOCK STRUCTURE IS CONSERVED ACROSS MAMMALS

***Guryev V.*[*1], *Smits B.M.G.*[1], *van de Belt J.*[1], *Verheul M.*[1], *Hubner N.*[2], *Cuppen E.*[1]**
[1] Hubrecht Laboratorium, Uppsalalaan 8, 3584CT, Utrecht, The Netherlands; [2] Max-Delbruck-Center for Molecular Medicine (MDC), 13092, Berlin-Buch, Germany
[*] Corresponding author: e-mail: guryev@niob.knaw.nl

Genetic variation in genomes is organized in haplotype blocks and species-specific block structure is defined by differential contribution of population history effects in combination with mutation and recombination events. Haplotype maps characterize the common patterns of linkage disequilibrium in populations and have important applications in the design and interpretation of genetic experiments. Although evolutionary processes are known to drive the selection of individual polymorphisms, their effect on haplotype block structure dynamics has not been shown.

Here, we present a high-resolution haplotype map for a 5 Mb genomic region in the rat and compare it with the orthologous human and mouse segments. Although the size and fine structure of haplotype blocks are species-dependent, there is a significant interspecies overlap in structure and a tendency for blocks to encompass complete genes. Extending these findings to the complete human genome using haplotype map phase I data reveals that linkage disequilibrium values are significantly higher for equally spaced positions in genic regions, including promoters, as compared to intergenic regions, indicating that a selective mechanism exists to maintain combinations of alleles within potentially interacting coding and regulatory regions.

The observed haplotype block conservation has several implications for experimental genetic approaches. First, it may complicate genetic studies, as identification of a single causal polymorphism underlying a QTL may turn out to be unfeasible and focus may switch to combinations of tightly linked alleles. Secondly, multi-specific approaches, using several model organisms at a time for narrowing down the QTLs region, may be less effective than anticipated. On the other hand, maps of conserved haplotype structures could point towards genomic segments that are under clear selection pressure in mammalian species and may allow for the identification of functional genomic elements, including important promoter and enhancer regions.

USING BIOLOGICAL KNOWLEDGE IN COMPUTATIONAL METHODS TO DISCOVER MECHANISMS OF TRANSCRIPTION REGULATION

Martin D.[1], Portales-Casamar E.[1], Kirov S.[2], Lim J.[1], Brumm J.[1], Snoddy J.[2], Wasserman W.W.[*1]

[1] Centre for Molecular Medicine and Therapeutics, CFRI, University of British Columbia, Vancouver, BC, CANADA; [2] Biomedical Informatics Dept, Vanderbilt University School of Medicine, Nashville, TN, USA

[*] Corresponding author: e-mail: wyeth@cmmt.ubc.ca

An essential mechanism for controlling gene expression is the binding of transcription factors (TFs) to specific regulatory elements (transcription factor binding sites; TFBSs). Pattern discovery algorithms (e.g. Gibbs Sampling) identify over-represented patterns consistent with TFBSs in sets of co-expressed genes.

We hypothesize that the discovery of regulatory mechanisms will be greatly enhanced by incorporating biological knowledge from in vivo, ex vivo and/or in vitro experiments into the pattern discovery process. By incorporating existing knowledge about key TFs and/or laboratory-defined TFBSs or regulatory regions, the pattern discovery process will be constrained, greatly reducing the amount of background noise that must be overcome in producing relevant predictions. In order to ascertain the impact of prior knowledge on the pattern discovery process, we have implemented a new Gibbs sampling procedure. Our algorithm is multi-phased, aiming at discovering, in sets of co-expressed genes, both regulatory regions and the over-represented motifs in those regions, the use of prior knowledge influencing both steps. Preliminary results will be presented.

A pre-requisite for this investigation is to compile enough regulatory information on co-expressed genes. In order to facilitate the collection of new data by the whole scientific community, we have developed the PAZAR database to store information related to gene regulation. PAZAR benefits experimentalists by providing a more efficient means to share TFBS information, accelerating experimental design. For computational biologists, the shared resource provides a richer range of reference data for the assessment of predictive algorithms.

The integration of the regulatory sequence annotations from PAZAR with the knowledge-based CRM discovery software offers promise for deciphering human gene regulatory mechanisms.

MOLECULAR NETWORKS IN MAMMALS: EXTRACTION FROM LITERATURE AND MICROARRAY ANALYSIS

Mazo I.*[*]*, Sivachenko A., Yuryev A., Daraselia N.
Ariadne, Rockville MD, 20850, USA
[*] Corresponding author: e-mail: mazo@ariadnegenomics.com

Using the proprietary high-content linguistics tool MedScan we compiled a comprehensive database of molecular networks by extracting the information from scientific literature. MedScan is capable of extracting functional associations between proteins, small molecules, and pathways, recognizes types of regulatory mechanisms involved, effects of regulation and experimental conditions. The resulting database stores 900,000 relationships between mammalian proteins and chemicals including facts about protein interactions, promoter binding, modification and protein regulation. Different approaches towards reconstructing individual pathways or cascades from this database and reconstructing networks from microarray data will be described. Our visualization software is capable of systematically mining this database for small network motifs that are robust in regard to the effects induced at the gene expression levels. We have also developed a Bayesian framework for integration of microarray data and binary gene-to-gene regulatory relationships. The approach allows the reduction of expression pattern complexity and finds the minimal set of regulatory proteins that are responsible for differential expression of other genes.

THE FEATURES OF STRUCTURAL DYNAMICS OF DIFFERENT TUBULIN SUBUNITS

Nyporko A.Yu.**, Blume Ya.B.
Institute of Cell Biology and Genetic Engineering of NAS of Ukraine, Kiev, Ukraine
e-mail: dfnalex@univ.kiev.ua

The molecular dynamics of α-, β- and γ-subunits of animal (*Sus scrofa* L.) and plant (*Arabidopsis thaliana* L.) tubulins – the main components of microtubules – was computed using molecular mechanics software *GROMACS* and their structural alterations were evaluated using *do_dssp* and *g_rms* software modules. The dynamics calculations were carried out for different time intervals (from 2 ps to 10 ns). It was revealed, that the secondary structure of all studied proteins is characterized by the evident metastability in time – a significant number of residues from helical and sheet structural elements drifts into unordered structures and goes back. Some of the structural elements, sheets S3 and S7 in particular, don't have stable cores. The observed effect can be one of the key factors providing a known phenomenon of dynamic instability of microtubules in living cell. However, the level of metastability of plant as well as animal γ-tubulins is less in comparison with stability level of other tubulin subunits – its value is 46 % against 50 % for β-tubulins and 54 % for α-tubulins. That can be conditioned by features of cell functions of γ-tubulins as the main component of MTOCs (microtubule organization centers), which are more stable formations in comparison with microtubules. The comparison of obtained results with our data of comparative alignment of the all known entire sequences within α-, β- and γ-tubulin families has shown that the representatives of most conservative tubulin family (α-tubulins) are characterized the most instability of secondary structure. And vice versa, the most variable by the sequences γ-family has the most stability structure in time. One should mention that structural drifting of all tubulin subunits in short time intervals (~ 100 ps) well reflects the structural alterations that are being observed during long time interval (10 ns).

RegulonDB: GOING BEYOND TRANSCRIPTIONAL REGULATION

Peñaloza-Spínola M.I., Peralta-Gil M.*, Gama-Castro S.*, Contreras-Moreira B., Santos-Zavaleta A., Martínez-Flores I., Collado-Vides J.
Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México (Center of Genomic Sciences, UNAM). Cuernavaca, Morelos; México
* Corresponding authors: martin@ccg.unam.mx; sgama@ccg.unam.mx

The most complete collection of literature-based information on the transcriptional regulatory network of *Escherichia coli* is available in the specialized database RegulonDB. This resource comprises the network elements (genes, operons, promoters, sigma factors, transcriptional factors, regulons and terminators) and the type and number of interactions. The current curated network has prompted the development of integrated models (Martinez-Antonio *et al*., 2003, Biotechnol Bioeng 84(7):743–9) and has also been used to drive biological network analyses (Shen-Orr *et al*., 2002, Nat Genet (1):64–8; Balazsi *et al*., 2005, Proc. Natl. Acad. 102(22):7841–6). Advancing in the knowledge of the cell networks, the RegulonDB team is now curating the information on other kind of regulatory mechanisms present in *E. coli* such as: post-transcriptional, translational, post-translational, epigenetic. The model strategy and the type of objects to be considered for the integration of these mechanisms with the transcriptional information is still under evaluation. Here we present two examples of main physiological processes, as the first approaches of this new integrated model. These examples involve transcriptional regulation in response to metabolism modifications that imply important physiological changes: oxygen starvation (anaerobiosis) and the metabolism of carbon sources changes. Our main goal is to complete the cell regulatory networks knowledge and its understanding.

RegulonDB. THE MOST IMPORTANT DATABASE IN TRANSCRIPTIONAL REGULATORY NETWORK, OPERON ORGANIZATION, AND GROWTH CONDITIONS OF *ESCHERICHIA COLI* K12

Santos-Zavaleta A., Salgado H.*[*], Gama-Castro S., Peralta-Gil M., Contreras-Moreira B., Peñaloza-Spínola M.I., Collado-Vides J.
Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México (Center of Genomic Sciences, UNAM). Cuernavaca, Morelos; México
[*] Corresponding author: e-mail: heladia@ccg.unam.mx

RegulonDB is the most important specialized database in transcriptional regulation, operon organization and regulatory network in *Escherichia coli* K12. Currently, it is the largest electronically-encoded database of the regulatory network of any free-living organism. Continuous curation of original scientific literature provides the evidence behind every single object and feature, and it is complemented with computational predictions of promoters, transcription units and DNA binding sites for regulatory proteins across the complete genome. The complex biology of regulation is simplified in a navigation scheme based on three major streams: genes, operons and regulons. Regulatory knowledge is directly available in every navigation step. Displays combine graphic and textual information and are organized allowing different levels of detail and biological context.

Based on the compiled data, we have development some tools which facilitate the use and analysis of its content. These have been associated to RegulonDB such as: GetTools (http://www.ccg.unam.mx/Computational_Genomics/GETools/) to facilitate the comparison with microarray experiments and links to the Regulatory Sequence Analysis (http://embnet.cifn.unam.mx/rsa-tools/), the Genome Browser, a new network graphic display or Network Tool, as well as links to Nebulon, a tool to predict groups of functionally related genes.

We have started expanding the curation beyond transcriptional regulation. In addition, we have been curating different physiological systems, information which will be soon available in RegulonDB. Our main objective is to achieve a complete view of the cell regulatory networks knowledge.

USE MOLECULAR MARKERS FOR DIFFERENTIATION POPULATIONS OF *STIPA CAPILLATA* GROWING IN THE REGIONS WITH HIGH CHRONICAL DOSES OF γ-RADIATION

Sarsenbaev K.N.
Institute of radiation safety and ecology, Kurchatov, Kazakhstan
e-mail: kanatsarsenbaev@hotmail.com

The aim of our study is to detect genetical diversity of 36 populations of Stipa capillata, based on their individual RAPD, MGH, α- and β-amylase, phosphatase, peroxidase, nonspecific esterase, polypeptides, soluble proteins. This patterns have been use to provide more objective analysis genotype and genetic relationships between radiated and non-radiated by γ-radiation population.

DNA from investigated populations was analyzed by RAPD-method with 10-15 nucleotide primers. Dendograms was constracted based on the similarity matrix data by applying the unweighted pair group method with arithmetic averages (UPGMA) cluster analysis. Proteins and enzymes by PAG-electrophoresis.

Among the examined 15 primers only four showed polymorphic pattern. The number of amplicons obtained by primers varied from 9 to 39. High level of radiation change compound complex of studied enzymes and proteins, but deference between populations growing on uncontaminated areas on the level of used markers very low. It was proposed that growing Stipa plants during 40 years on the contaminated areas leads to appearance new genotypes.

UNEQUALLY SPACED SAMPLING MAPPING OF QTL

Shaoqing Huang**, *Yini Cui
College of Life Science and Biotechnology, Shanghai Jiao Tong University, Shanghai, China
* Corresponding author: e-mail: ShaoqingHuang@gmail.com

We described a general statistical strategy for accommodating unequally spaced quantitative trait data in QTL mapping. Our heuristic is based on the notion that, in constructing the residual covariance matrices, by inferring the underlying relation of QTLs within inbred strains, the missing phenotype data could be covered when computing weights, which are assigned to the simulated genotypes to capture information in the phenotype data. Our method is based on the premises that there is the same residual variance for the trait at each time and covariance stationarity. Thus we developed a simple Legendre Polynomial algorithm to implement QTL analysis and applied order one Auto-Regression model to estimate the genetic covariance functions. The power of unequally spaced sampling mapping is demonstrated by an example of a forest tree, in which one QTL affecting stem growth processes is detected by functional mapping. We randomly deleted *X* years' data sets from the altogether eleven-year diameter data and input them to our computer programs. The results showed that when *X* is between one to four, the unequally spaced sampling mapping result is almost identical to the results computed from the whole data sets, with exception for some minor differences. Moreover, there is no evidence that the differences are specific for deletion of certain data sets. Furthermore, we advanced our work to disequilibrium sampling mapping that allows each individual having its own number of data sets.

LABORATORY INFORMATION MANAGEMENT SYSTEM FOR MEMBRANE PROTEIN STRUCTURE INITIATIVE

***Troshin P.V.*[*1], *Papiz M.Z.*[1], *Prince S.M.*[2], *Daniel E.J.*[3], *Morris C.*[*3], *Griffiths S.L.*[4], *Diprose J.M.*[5], *Pilicheva K.*[5], *v. Niekerk J.*[6], *Pajon A.*[7]**

[1*] Membrane Protein Structure Initiative, CCLRC Daresbury Laboratory, UK; (Questions relate to MPSI) - p.v.troshin@dl.ac.uk

[2] Faculty of Life Sciences, University of Manchester, Sackville St., Manchester, UK; [3] Protein information management system, CCLRC Daresbury Laboratory, UK; (Questions relate to PIMS) - c.morris.dl.ac.uk

[4] York Structural Biology Laboratory, University of York, UK; [5] Oxford Protein Production Facility, Division of Structural Biology, University of Oxford, UK; [6] Scottish Structural Proteomics Facility, University of Dundee, UK; [7] European Bioinformatics Institute, Hinxton, UK.

[*] Corresponding author

The advent of structural genomics in the UK with the newly established BBSRC MPSI consortium, demonstrates a growing need for a sophisticated LIMS suitable for academic environments.

The Membrane Protein Structure Initiative (MPSI) is a consortium formed by highly experienced and successful membrane protein research groups in order to use their collective experience and techniques to develop strategies for high throughput (HTP) expression, purification, characterisation, crystallisation and structure determination of selected integral membrane proteins. Shared data and results are entered in a common LIMS database to be made freely available to all participating labs through the web site (www.mpsi.ac.uk).

Establishing new, large, geographically distributed consortia such as MPSI introduces novel communications and data exchange problems, which, together with the wide spread of high throughput techniques, dictate a critical need for innovative information management and knowledge discovery tools to sift through these vast volumes of heterogeneous data and analysis tools.

As part of the LIMS, the BBSRC has funded the Protein Information Management System (PIMS) project to meet this need. PIMS is an adaptable and scalable information management system. The project is based on the Protein Production Data model, which was developed jointly by the EBI and CCPN. The system automates a range of tasks from protein sample preparation from selection, cloning, expression, purification, characterisation and, where applicable, crystallisation through to automated data deposition in the PDB.

The project itself provides a mechanism for the incremental development and deployment of such a system, by drawing together all the interested parties to work under a single framework which is committed to meeting the needs of the MPSI consortia and other contributing groups. An initial version with limited functionality is available for download from the project website, www.pims-lims.org .

Indexes

AUTHOR INDEX

KEY WORDS

Научное издание

Труды пятой международной конференции
“Биоинформатика регуляции и структуры генома”
Т. 3
на английском языке

Proceedings of the Fifth International Conference
on Bioinformatics of Genome Regulation and Structure
V. 3

Отредактировано и подготовлено к печати
в редакционно-издательском отделе ИЦиГ СО РАН
Редакторы: А.А. Ончукова, И.Ю. Ануфриева
Дизайн А.В. Харкевич
Компьютерная графика: А.В. Харкевич, К.В. Гунбин, Т.Б. Коняхина
Компьютерная верстка: А.В. Харкевич, К.В. Гунбин, Н.С. Глазкова

Подписано к печати 21.06.2006 г.
Формат бумаги 70×108 1/16. Печ.л. 28,9. Уч.-изд.л. 31,5
Тираж 250. Заказ 271

Институт цитологии и генетики СО РАН
630090, Новосибирск, пр. акад. М.А. Лаврентьева, 10
Отпечатано в типографии Издательства СО РАН
630090, Новосибирск, Морской пр., 2