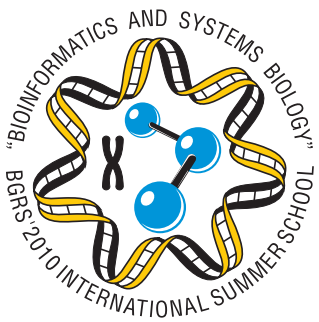
RUSSIAN ACADEMY OF SCIENCES
SIBERIAN BRANCH

INSTITUTE OF CYTOLOGY AND GENETICS

The Young Scientists School

"BIOINFORMATICS AND SYSTEMS BIOLOGY"



Program
&
Abstracts

Novosibirsk, Russia
June 28–29, 2010

INTERNATIONAL PROGRAM COMMITTEE

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

Georges St. Laurent, St. Laurent University, Providence, USA, (Co-chairman)

Dmitry A. Afonnikov, Institute of Cytology and Genetics, Novosibirsk, Russia (Co-chairman)

Igor Rogozin, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, USA

Fazel Famili, University of Ottawa, IIT/ITI - National Research Council Canada, Ottawa, Canada

Henrik Jonsson, Lund University, Sweden

Pawel Krupinski, Lund University, Sweden

Lavrentiev Mikhail, Novosibirsk State University, Novosibirsk

Alexis Ivanov, Institute of Biomedical Chemistry RAMS, Moscow, Russia

Daniil G. Naumoff, State Institute for Genetics and Selection of Industrial Microorganisms, Moscow, Russia

Olga Krebs, Heidelberg Institute for Theoretical Studies, Germany

Thomas Liehr, Jena University Hospital, Institute of Human Genetics and Anthropology, Germany

LOCAL ORGANIZING COMMITTEE

Victoria Mironova, Institute of Cytology and Genetics, Novosibirsk, Russia (Chairperson)

Ilya Akberdin, Institute of Cytology and Genetics, Novosibirsk, Russia

Luba Glebova, Institute of Cytology and Genetics, Novosibirsk, Russia

Nadezda Glebova, Institute of Cytology and Genetics, Novosibirsk, Russia

Tatyana Karamysheva, Institute of Cytology and Genetics, Novosibirsk, Russia

Andrey Kharkevich, Institute of Cytology and Genetics, Novosibirsk, Russia

Galina Kiseleva, Institute of Cytology and Genetics, Novosibirsk, Russia

Svetlana Zubova, Institute of Cytology and Genetics, Novosibirsk, Russia

Ree Maksim, Institute of Cytology and Genetics, Novosibirsk, Russia

Galina Orlova, Institute of Cytology and Genetics, Novosibirsk, Russia

Organizers

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

Siberian Branch of the Russian Academy of Sciences

The Vavilov Society of Geneticists and Breeders

Laboratory of Theoretical Genetics

Novosibirsk State University

Chair of Information Biology

PBSoft Ltd.

Sponsors

Federal Agency for Science and Innovation

Russian Foundation for Basic Research

Contents

THE YOUNG SCIENTISTS SCHOOL PROGRAM

The young scientists school sessions is held in the Institute of Cytology and Genetics, conference hall

Chairs of the young scientists school: *Dr. Dmitry A. Afonnikov*
Prof. Georges St. Laurent
Acad. Nikolay A. Kolchanov

June, 28

8.30–9.30 registration

9.30–10.15 Welcome by academician of RAS, **Nikolay Kolchanov**
Prof., St. Laurent University, USA, **Georges St. Laurent**
Dr., Institute of Cytology and Genetics, **Dmitry A. Afonnikov**

Morning session (9.30–13.00)

9.30–10.15 Prof. **Thomas Liehr** (Jena University Hospital, Institute of Human Genetics and Anthropology, Germany)
Small supernumerary marker chromosomes in human

11.00–11.30 coffee break

11.30–12.15 Prof. **Henrik Jönsson** (Lund University, Sweden)
Combining morphodynamics, hormone signaling, and cell differentiation in computational models

12.15–13.00 Dr. **Pawel Krupinski** (Lund University, Sweden)
Biomechanics of cells and their interactions - models for plants and animals

Evening session (14.00–18.00)

14.00–14.45 Prof. **Georges St. Laurent III** (St. Laurent Institute, Providence, USA)
Computational mechanisms and information coding by the non-coding transcriptome

14.45–15.30 Dr. **Olga Krebs** (Heidelberg Institute for Theoretical Studies, Germany)
Systems biology requirements for standardization and integration of wet and dry laboratory data

Evolutionary genomics of eukaryotes

15.30–16.00 coffee break

Young scientists' reports

16.00–16.15 **Berillo O.A., Khailenko V.A.** (Kazakh National University named al-Farabi, Almaty, Kazakhstan)
Peculiarities of interaction miRNA with mRNA of some oncogenes

16.15–16.30 **Galachyants Yu.P. (**Limnological Institute, Irkutsk, Russia)
Metagenomics analysis of the bacterial community associated with diatom alga Synedra acus

16.30–16.45 **Khlopova N**. (RSAU-MTAA named after K.A.Timiryazev, Moscow, Russia)
Variable part of gene expression profiles in liver and kidney of pigs

16.45–17.00 **Gurianova V.** (Bogomoletz Institute of Physiology, Kiev, Ukraine)
Integral evaluation of profile of natriuretic peptides system's mRNA expression in cultured cardiomyocytes during anoxia-reoxygenation

17.00–17.15 Coffee break

17.15–17.30 **Arakelyan A.** (Institute for Informatics and Automation Problems, Yerevan, Armenia)
Algorithmic analysis of functional pathways affected in post-traumatic stress disorder

17.30–17.45 **Sirotkina S.M.** (Tyumen State University, Tyumen, Russia)
Application of DNA sequencing and detection of levels expression epidermal growth factor receptor (EGFR) in cancer practice

17.45–18.00 **Shadrin A.A.** (Design Technological Institute of Digital Techniques SB RAS, Novosibirsk, Russia)
Comparison of methods for reconstruction of models for gene expression regulation

June, 29

Morning session (9.30–13.00)

9.30–10.15 Prof. **Lavrentiev Mikhail** (Novosibirsk State University, Novosibirsk)
Modern hardware architectures and acceleration of solution to some problems in biology

10.15–11.00 Prof. **Fazel Famili** (University of Ottawa, IIT/ITI - National Research Council Canada, Ottawa, Canada)
The real impact of knowledge discovery in bioinformatics

11.00–11.30 coffee break

11.30–12.15 Dr. **Daniil G. Naumoff** (State Institute for Genetics and Selection of Industrial Microorganisms, Moscow, Russia)
Bioinformatic analysis of protein families

12.15–13.00 Prof. **Alexis Ivanov** (Institute of Biomedical Chemistry RAMS, Moscow, Russia)
Intermolecular interactions: 3D computer simulations and SPR biosensor analysis

Evening session (14.00–18.00)

Young scientists' reports

14.00–14.15 **Medvedeva I.V.** (Institute of Cytology and Genetics, Novosibirsk, Russia)
Computer system SitEx for analyzing protein functional sites in eukaryotic gene structure

14.15–14.30 **Genaev M.A**. (Institute of Cytology and Genetics, Novosibirsk, Russia)
BioinfoWF — web services and workflow management for bioinformatics analysis

14.30–14.45 **Uroshlev L.A.** (Moscow State Forest University, Mytischi, Russia)
GPGPU-computing for prediction of small ligand binding sites in proteins

14.45–15.00 **Novoseletsky V.N.** (Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Moscow, Russia)
Influence of hydrophobicity of beta-blockers on their binding affinity

15.00–15.15 **Mikhailova E.V.** (Institute of Cytology, Saint-Petersburg, Russia)
Prion-associated proteinsin yeast: comparative analysis of yeast strains, distinguished by their prion content

15.15–15.30 **Chernobrovkin A.L.** (Institute of Biomedical Chemistry RAMS, Moscow, Russia; Cryptome Research Ltd, Moscow, Russia)
Knowledge-based refinement in mass-spectrometric identification of proteins

15.30–16.00 coffee break

16.00–16.15 **Stepuschenko O.O.** (Kazan State University, Department of Genetics, Kazan, Russia)
Sequence analysis of COG3868 and COG2342 families

16.15–16.30 **Asheulov A.S.** (Al-Farabi Kazakh National University, Almaty, Kazakhstan)
Exon-Intron structure of first chromosome Monodelphis domestica

16.30–16.45 **Orlenko A.** (Novosibirsk State University, Novosibirsk, Russia; Institute of Cytology and Genetics, Novosibirsk, Russia)
The correspondence between 5'-untranslated regions of yeast's genes with their expression efficiency.

16.45–17.00 **Bukin Yu.S.** (Limnological Institute SB RAS, Irkutsk, Russia)
Model for simulation gene flow for species with one-dimensional area

17.00–17.15 **Kuznetsova E.** (Siberian Institute of Plant Physiology and Biochemistry, Irkutsk, Russia)
Dwarf form of Malus baccata (L.) Borkh: initial stage of the parapatric speciation?

17.15–17.30 **Doroshkov A.** (Institute of Cytology and Genetics, Novosibirsk, Russia)
Using the computer-based image processing technique in genetic analysis of leaf hairiness in wheat Triticum aestivum L.

18.00 **The close ceremony, award, summaries from chairpersons**
academician of the Russian academy of Science, **Nikolay Kolchanov**
Prof., St. Laurent University, USA, **Georges St. Laurent**
Dr., Institute of Cytology and Genetics, **Dmitry A. Afonnikov**

The abstracts of the lectures

at the young scientists school

Prof. Thomas Liehr

Jena University Hospital, Institute of Human Genetics and Anthropology, Germany

SMALL SUPERNUMERARY MARKER CHROMOSOMES IN HUMANS

Small supernumerary marker chromosomes (sSMC), defined as additional centric chromosome fragments too small to be identified or characterized unambiguously by banding cytogenetics alone, are present in 0.043% of newborn children. Several attempts have been made to correlate certain sSMC with a specific clinical picture, resulting in the description of several syndromes such as the Pallister Killian syndrome or Turner syndrome and others. However, most of the sSMC including minute-, ring-, inverted-duplication- as well as complex-rearranged chromosomes, have not yet been correlated with clinical syndromes, mostly due to problems in their comprehensive characterization. Breakpoint characteristics of sSMC are not well studied by now. Array-comparative genomic hybridization (aCGH) was done in 64 sSMC. A detailed analysis of overall 128 characterized breakpoints in non-heterochromatic chromosomal regions of sSMC was done that reveals breakpoint hot spots.

Prof. Olga Krebs

Heidelberg Institute for Theoretical Studies, Germany

SYSTEMS BIOLOGY REQUIREMENTS FOR STANDARDISATION AND INTEGRATION OF WET AND DRY LABORATORY DATA

Systems biology involves analyzing and predicting the behaviour of complex biological systems like cells or organisms. This requires qualitative information about the interplay of genes, proteins, chemical compounds, and biochemical reactions. A challenge in many Systems Biology projects is management of different data types and linking experimental to theoretical and modelling approaches. Data types range from high-throughput, top down measurements to individual parameter estimations in bottom-up approaches. Combining these data sets and linking them to kinetic models requires a good set of bioinformatics and systems biology tools.

SySMO DB (http://www.sysmo-db.org/) is a web based platform for finding, sharing and exchanging data, models and processes developed as a data integration platform across the SysMO consortium (Systems Biology for MicroOrganisms; http://www.sysmo.net/), but the principles and methods employed are equally applicable to other multisite Systems Biology projects. The support for data exchange is accomplished by implementing a "Just Enough Results Model" (JERM) approach, by which we define the least amount of information required to enable sharing and exchange. A JERM for any one type of data (i.e. microarray data, or metabolomic data) is the minimum data schema that SysMO projects agree to share.This session will focus on existing standards, databases

and tools that are relevant to systems biology and discuss a number of key issues affecting data integration and the challenges these pose to systems-level research.

Prof. Henrik Jonsson

Lund University, Sweden

COMBINING MORPHODYNAMICS, HORMONE SIGNALING, AND CELL DIFFERENTIATION IN COMPUTATIONAL MODELS

The mathematical decription of multicellular development needs to take into account additional mechanisms compared to a single cell systems biology description. Cells are growing and proliferating and molecular and mechanical signalling between cells are of importance. In this talk I will focus on these aspects when it comes to computational models and I will provide examples from our own research, mainly from plant development and quorum sensing within bacterial colonies. In particular I will focus on how cellular mechanisms lead to global behavior.

Dr. Pawel Krupinski

Lund University, Sweden

BIOMECHANICS OF CELLS AND THEIR INTERACTIONS — MODELS FOR PLANTS AND ANIMALS

The diversity of mechanical properties, forms and functions of biological tissues is enormous. How this diversity translates to and originates from mechanical behavior of individual cells is fundamental question of morphodynamics.

I will review mechanical, cellular models of living tissues and discuss their applicability to different tissue types. In these models the cell consists a basic building unit of the tissue, which properties and interactions with surroundings influence its fate in the course of the development.

I will present examples of finite element method modeling of a plant tissue in *Arabidopsis thaliana* meristem, as well as agent based modeling of early mammalian embryogenesis in mouse mouse embryo. I will show how mechanical aspects of morphogenesis in these models have to be coupled to biochemical interactions to give the picture of underlying processes consistent with experimental data.

Prof. Georges St. Laurent III

St. Laurent Institute, Providence, USA

COMPUTATIONAL MECHANISMS AND INFORMATION CODING BY THE NON-CODING TRANSCRIPTOME

The rapidly expanding field of non-coding genomics has illuminated a new horizon: the scope and complexity of animal transcriptomes. The FANTOM Consortium produced organism-wide cDNA libraries of mouse transcripts, including a surprising 23,000 non-coding RNAs (ncRNAs), compared to 20,929 conventional mRNAs found. FANTOM IV and ENCODE have since confirmed that a majority of transcripts in mouse are in fact noncoding. This emerging ncRNA landscape composes a computational matrix that functions to acquire, process, and distribute biological information.

ncRNAs could contribute to the success of the organism's information processing in several ways. First, ncRNAs would allow for efficient coupling of energy with information, wherein less energy is required to represent and process more information, condensed in analog and digital form, into smaller spatial and temporal domains, ideal for the environments found in neural tissues. Second, ncRNAs would permit the rapid acquisition of information from the environment, along with the rapid flexible processing and elimination of that information when it is no longer necessary. Third, ncRNAs would facilitate accelerated evolution of an organism's information content and functional computational systems. This emerging panorama might open new dimensions of cellular information processing.

The depth and complexity of the non-coding transcriptome in mammalian tissues provides a rich substrate for computational processing, coupling signal flow from the environment to evolutionarily coded analog and digital information elements within the transcriptome. We present a perspective of the non-coding transcriptome as a computational matrix, enhancing the information processing power of the cell, adding flexibility, rapid response, and fine tuning to critical pathways.

Prof. Lavrentiev Mikhail

Novosibirsk State University, Novosibirsk

MODERN HARDWARE ARCHITECTURES AND ACCELERATION OF SOLUTION TO SOME PROBLEMS IN BIOLOGY

Brief introduction into facilities of modern Graphic Processing Units (GPU) and Field Programmable Gates Arrays (FPGA) will be delivered. Use of GPU is rather simple for those who know C++ computer language, the so called CUDA environment (being designed to use GPU for applied calculations) is well described and easy to exploit. At

the price of slight code modification it is possible to reduce the required computer time from 14 days with PC to just 7-8 hours. FPGA could be regarded as modern template for special processor. Model evaluation shows that in case of use of FPGA based printed board computation time could be reduced to just 22 minutes.

Prof. Fazel Famili

University of Ottawa, IIT/ITI – National Research Council Canada, Ottawa, Canada

THE REAL IMPACT OF KNOWLEDGE DISCOVERY IN BIOINFORMATICS

Over twenty five years ago, a famous computer scientist and Nobel Laureate, by the name H.A. Simon, noted that the most fundamental consequence of the superabundance of huge volume of data and information created by the digital revolution had resulted in lack of attention to the real value of: (i) data due to lack of time, tools and expertise for proper knowledge discovery and (ii) information and knowledge that is discovered subsequently. Over the last 5-10 years, the exponential growth of databanks in life sciences has also created an opportunities to further expand research topics, such as bioinformatics, from a knowledge discovery point of view. An example is the development of scientific approaches to intelligently "mine" the huge databanks that complex systems rely on for their management. Of the more complex of all data domains are the life sciences where one tries to integrate and analyze large amounts of high-throughput omics data (such as genomics and proteomics) obtained from either single time point or time-series applications. Similar to many other domains, in life sciences various methods have been developed, and many knowledge discovery tools have been introduced. The objective of this lecture is to provide an overview of knowledge discovery in bioinformatics and explain the impact that knowledge discovery can make in the bioinformatics field.

Dr. Daniil G. Naumoff

State Institute for Genetics and Selection of Industrial Microorganisms, Moscow, Russia

BIOINFORMATIC ANALYSIS OF PROTEIN FAMILIES

About 30 millions of protein sequences are deposited into the databases. They represent at least 80 thousands of protein domain families. The vast majority of the proteins has not been and never will be characterized experimentally. So, we need to predict their functions based on the sequence analysis. The most powerful way is to extrapolate the available experimental data on individual proteins to their homologues (for example, within the protein family).

I am going to speak about protein databases, searching of homologous proteins, protein architecture (domain structure), multiple sequence alignment, phylogenetic tree

reconstruction, etc. I will present a complete procedure of a protein family analysis, viz. from database searching to visualization of the phylogenetic tree. The analysis can be started using any protein sequence as a query. The methods and programs suggested can be applied to any protein family but they would work more effectively with globular solving proteins.

Prof. Alexis Ivanov

Institute of Biomedical Chemistry RAMS, Moscow, Russia

INTERMOLECULAR INTERACTIONS: 3D COMPUTER SIMULATIONS AND SPR BIOSENSOR ANALYSIS

Various intermolecular interactions are the corner stone of the live organisms functioning. Therefore virtual and experimental investigations of intermolecular interactions are the one of fundamental tasks of modern system biology. The given lecture is addressed to young scientists to give them a short overview of two complementary approaches to the analysis of intermolecular interactions *in silico* and *in vitro*: (1) computer 3D analysis and simulation of molecular complexes; (2) experimental analysis of intermolecular interactions by using the optical biosensor on the effect of surface plasmon resonance.

The abstracts

of the Young scientists reports

ALGORITHMIC ANALYSIS OF FUNCTIONAL PATHWAYS AFFECTED IN POST-TRAUMATIC STRESS DISORDER

A.A. Arakelyan*, A.S. Boyajyan, L.A. Aslanyan
"Laboratory of Information Biology" Project of the Institute of Molecular Biology and Institute for Informatics and Automation Problems National Academy of Sciences of Republic of Armenia Yerevan, Armenia
e-mail: aarakelyan@sci.am
**Corresponding author*

Key words: *gene expression, pattern recognition, functional gene sets, post-traumatic stress disorder*

Motivation and Aim: Promising findings suggest that both environment and genetic factors are involved in post-traumatic stress disorder (PTSD) generation mechanisms, and expression alterations of genes involved in neuronal, endocrine, and immune function might be in a sufficient degree responsible for disease progression. In the present study we design and extend logic-combinatorial scheme to evaluate the functional gene sets affected in peripheral blood mononuclear cells of patients with posttraumatic stress disorder at both early and advanced stages of the disease.

Methods and Algorithms: We used dataset containing gene expression data from patients with post-traumatic stress disorder from Gene Expression Omnibus (GDS1020, http://www.ncbi.nlm.nih.gov/geo/). In general, dataset $\Im$ can be represented as $\Im = \{S_1, S_2, ..., S_n\}$, where S - a is column-vector of gene expression data for individual sample. Classes consist of: class 1 - $\{S_1, ..., S_k\}$ and class 2 $\{S_{k+1}, S_n\}$. The main idea is to find subset of genes that can maximally separate given classes by linear hyperplane. At the first step the search for the best single gene (1-classifier) separating classes with maximal accuracy is performed. If more than one gene is identified, the separation force is computed and gene with maximal score is selected. At the next step the gene pair consisting of 1-classifier and other gene with maximal separation force is selected. The separation force is calculated as a distance of class center and separation hyperplane. The iterations are performed until maximal classification accuracy is achieved. To speed up calculation processes the last 5% of classifiers with worst classification accuracy are removed at each step. The algorithm was implemented in Matlab R2008a (Mathworks Inc, USA).

Results: Using the proposed algorithm we identified set 63 genes that separate classes (patients with PTSD/healthy control) with 100% accuracy. Further analysis of set showed that in PTSD upregulated genes are involved in inflammatory, cell sycle/proliferation, signal transduction pathways, whereas downregulated genes are involved regulation of transcription, mitochondrial function and biosynthetic processes.

Conclusion: Our results give biological insights into molecular processes occurring in PTSD. In addition, extended classification algorithms are designed to analyze high dimension low sample size data.

Availability: Matlab m-files available on request from the authors.

Acknowledgements: We gratefully acknowledge the Segman R.H., et al and Renthal W., et al for making their data publicly available. We also thank L. Nersisyan for her assistance.

EXON-INTRON STRUCTURE OF FIRST CHROMOSOME MONODELPHIS DOMESTICA

A.S. Asheulov*
Kazakh National University named after al-Farabi, Almaty, Kazakhstan
e-mail: laxanier@mail.ru
**Corresponding author*

Keywords: *exon, intron, genome, computing analysis.*

Motivation and Aim: Introns perform many functions in genes, and the analysis of their properties is necessary for the further understanding of their biological role. Variability ofexon and intron lengths is rather great in genes of eukaryotic organisms. In genomes of the human, nematode, arabidopsis and rice the links between exon and intron length, and the sum of exon lengths and intron number in genes have been established [1, 2]. It is obviously important to clarify if there are such links in genes of completely sequencing genomes of other eukaryotic organisms.

Methods and Algorithms: The genes of chromosome 1 Monodephis domestica were arranged in samplings with 1, 2, 3, 4, 5, 6-9, 10-14, 15 and more introns in a gene.

*Result*s: 4840 genes have been analysed. The average length of genes for groups with the maximum density made 1304 nucleotides, in groups with average density 866 nucleotides, and in groups with the minimum density equaled 961 nucleotide. Also it is necessary to notice that at quantity lengthening introns in a gene average relation C/G – A/T decreased. If to take group with 1-2 introns that this size makes 0,499 and if to take this size with 15 and more introns that it already 0,069. Average GC the maintenance in a chromosome 1 has made 41 %. It is necessary to notice that in group with smaller quantity introns size C/G-A/T it is essential more than in groups with a considerable quantity introns. This phenomenon is marked in all groups of sample with the minimum, average and maximum density. Than more long the sum of lengths introns in a gene that best effect. Also the change size gidropatichnosti in genes in groups with 1-2, 3-5,intron almost identical also equals 0,2. In groups with 6-9 introns the size fluctuates from 0,1-0,3. In groups from 10-14 introns and 15 and more changes in the field of 0,09-0,16. This data can testify to a different role introns in change gidropatichnosti genes with various density. It is necessary to add that the big role in change nucleotides structure introns bring in genes with the minimum density in groups with 3-5 and 6-9 introns in a gene, in genes with about average density in groups with 3-5 and 15 and more introns in a gene, in genes with the maximum density in groups with 1-2, 6-9 and also 15 and more intron in a gene.

References:

1. S.Atambaeva, V.Khailenko, A.Ivashchenko (2008) Changes of introns and exons length in genes of arabidopsis, rice, nematode and human, *Mol. Biol.* (Russ), **42:** 1-10.
2. A.Ivashchenko, S.Atambaeva (2004) Variation in lengths of introns and exons in genes of he *Arabidopsis thaliana* nuclear genome, *Russ. Journ. Genetics*, **40:** 1179-1181.

PECULIARITIES OF INTERACTION miRNA WITH mRNA OF SOME ONCOGENES

O.A. Berillo, A.S. Isabekova, V.A. Khailenko, A.T. Ivachshenko*
Kazakh National University named al-Farabi, Almaty, Kazakhstan
e-mail: a_ivashchenko@mail.ru
**Corresponding author*

Key words: *2D mRNA, miRNA, oncogenes, exon, intron, 3'UTR*

Motivation and Aim: The majority of miRNAs regulate expression of genes in eukaryotic cells at mRNA translation stage. miRNAs promote splitting of mRNA on fragments, reduce speed of translation, or activate translation process considerably [1, 2]. Usually miRNAs co-operate with 3'UTR and 5'UTR. However their interaction with introns in pre-mRNA and with exons of mRNA is revealed. We studied peculiarities of miRNA interaction with two-dimensional (2D) structure of mRNA.

Methods and Algorithms: 2D structure of both pre-mRNA and mRNA was built under the program UNAFold.3.7 (http://dinamelt.bioinfo.rpi.edu). Interaction sites of 2D mRNA with miRNAs were extracted from microRNA.org (http://www.microrna.org). Interaction of miRNAs with mRNA of *BAX*, *FASLG*, *FTL*, *GAST*, *TNFRSF17* oncogenes has been studied.

Result: Principal causes of cancer development are modification of gene expression, also at level of translation processes [1]. It was revealed, that all miRNAs co-operate with mRNA in sites which have unpaired nucleotides. On the average the number of unpaired nucleotides in sites of 2D structures of mRNA, connecting miRNAs, makes 35-60% out of the number of pairs formed at of miRNAs with mRNA interaction. It provides stronger linkage of miRNAs with mRNA. In some cases miRNAs co-operate with both complement sites of 2D mRNA that increases significant change 2D structures of mRNA. In mRNA of *BAX*, *FASLG*, *FTL*, *GAST* and *TNFRSF17* oncogenes there are sites connected with 10, 23, 3, 3 and 5 miRNAs respectively. Hence, the probability of updating of gene expression increases accordingly. mRNA of *BAX* and *FTL* genes have miRNAs linkage sites in exons. Six miRNAs contact with exon-6 and two miRNAs contact with exon-7 of *BAX* mRNA. mRNA of *FTL* gene contains six sites of miRNA linkage in exon-3 and 11 sites in exon-4. In exon-3 these sites are located in the site adjoining exon-4. Such features of interaction of miRNAs with mRNA of these genes can cause translation of truncated proteins. It means that to proteins synthesised on mRNA and received as a result of alternative splicing, proteins are added as products of alternative translation. Some miRNAs can influence splicing as they contact intron near to splicing sites. So, 7 miRNAs contact with intron-3 at the site splicing exon-3/intron-3 and intron-3/exon-4 of *TNFRSF17* pre-mRNA. In some genes the sites of interaction mRNA with miRNAs revealed the increase of CUG triplets which are specific for the restriction RNA.

Conclusion: Translation of mRNA of *BAX*, *FASLG*, *FTL*, *GAST* and *TNFRSF17* oncogenes can be essentially modified by molecules of miRNAs, which interact with 3'UTR, exons and introns. The obtained data promote the creation of oncodiagnostic and oncotherapeutic methods.

References:

1. I. L.de Silanes et al. (2007) Aberrant regulation of messenger RNA 3'-untranslated region in human cancer, *Cellular Oncology*, **29**: 1–17.
2. T.H.Beilharz et al. (2009) microRNA-mediated messenger RNA deadenylation contributes to translational repression in mammalian cells, *PLoS One*, **4**: e6783.

MODEL FOR SIMULATION GENE FLOW FOR SPECIES WITH ONE-DIMENSIONAL AREA

Yu.S. Bukin*
Limnological Institute SB RAS, Irkutsk, Russia
e-mail: bukiyura@mail.ru
**Corresponding author*

Key words: *populations, gene flow, gene drift, migration, population size, geographical* barrier.

Motivation and Aim: The study population structure of modern species using molecular genetic data is of great interest for scientist which connected the fact that the population represents a structural unit of microevolution, in which changes under the influence of gene drift and natural selection. The using genetic data, in particular molecular characteristics of organisms can most accurately define the extent of intraspecies genetic variability and geographic boundaries of populations. Often naturally species have ribbon area of habitats. If the time spent by the organism to cross intersection of the ribbon area is less than or comparable with the duration of a generation, all the processes associated with the formation of genetic diversity will be determined by the flow of genetic information along the area. Such area of habitats, in terms of genetics, can be considered relatively one-dimensional. The boundaries of the populations on the one-dimensional area is areas where suspended or significantly slows down the flow of genes along the area.

Methods and Algorithms: In our work, we offer a program that implements a model of population dynamics on the one-dimensional area of habitat. The main characteristic of area in model is the maximum allowable density of organisms at each point. User before the numerical experiment divides the one-dimensional area into segments in each of which specifies its maximum allowable density of organisms. Each individual in the model is given its own "nucleotide sequence", which passed from parent to offspring with mutation probability. At the end of each simulation researcher can compare the molecular data model with data obtained in the study of natural populations. On the basis of comparison model and real data it is possible to do conclusions about the applicability of the hypothesis of formation of population structure of species and genetic polymorphism. The algorithm of the program and the modeling method is based on an individually-based approach proposed in the papers [2,3] and then used by us in previous studies [1,4].

References:

1. Bukin Ju.S., Pudovkina T.A., Sherbakov D.Ju., Sitnikova T.Ya., 2007. Genetic Flows in a Structured One-Dimensional Population: Simulation and Real Data on Baikalian Polychaetes M. Godlewskii. // In Silico Biology; v. 7(3), pp. 277-284.
2. Dieckmann U., Doebeli M., 1999. On the origin of species by sympatric speciation. // Nature vol. 400, pp. 354-357.
3. Doebeli M., Dieckmann U., 2000. Evolutionary branching and sympatric speciation caused by different types of ecological interactions. // Am. Nat. vol. 156, pp. 77-101.
4. Semovski S.V., Bukin Y.S., Sherbakov D.Y., 2003. Speciation and neutral molecular evolution in one-dimensional closed population. // International journal of modern physics, vol. 14, pp. 973-983.

KNOWLEDGE-BASED REFINEMENT IN MASS-SPECTROMETRIC IDENTIFICATION OF PROTEINS

Chernobrovkin A.L.*[1,2], Lisitsa A.V.[1], Moshkovskii S.A.[1], Zgoda V.G[1], Archakov A.I.[1]
[1]*Institute of Biomedical Chemistry RAMS, Moscow, Russia*
[2]*Cryptome Research Ltd, Moscow, Russia*
e-mail: chernobrovkin@gmail.com
[*]*Corresponding author*

Key words: *protein identification, database searching, single amino acid polymorphisms*

Motivation and Aim: Database search is currently the dominant method in proteomic data analysis for the inference of proteins identification from tandem mass spectra. In high-throughput proteomics experiments with complex protein mixtures researchers must take a choice between the non-redundant databases, and large databases like EST that lead to a lot of false positive results. A possible decision is the use of iterative schemes in which non-redundant sequence database is used for protein identification at the first iteration and the refinement on previously identified proteins is performed at the second iteration.

Methods and Algorithms: Protein identification database PRIDE was used to download the mass-spectrometric data in mzData format from 3 experiments on the human protein identification. Mass-spectra were searched against UniProtKB/SwissProt database using cluster version of Mascot (Matrixscience Ltd.). UniProt knowledgebase was use to retrieve annotated single amino acids polymorphisms (SAPs) for the identified proteins (identification significance level was 0.05). For each experiment the new database was created by the addition of SAP-containing variants of previously identified proteins. The second iteration of database search was performed using new databases that allowed identifying of SAP-containing variants of the proteins.

Results: The database search of MS data of 3 experiments against the UniProtKB/SwissProt database results in 138 proteins comprising 1450 SAPs (~10 SAPs per protein) according to the UniProtKB annotations. The second iteration of database search was performed using individually created database for each experiment. Analysis of the Mascot search reports allows us to reveal 4 events of single amino acid polymorphism. One event corresponded to the residue change from Aspartate (D) to Clycine (G) in position 1073 was observed in complement protein C4 (CO4B HUMAN). The substitution of Threonine (T) with Methionine (M) in angiotensinogen (ANGT HUMAN) was observed in two experiments. In contrast to previously mentioned SAPs, designated by UniProtKB as "Polymorphism", T599R mutation in B-Raf protooncogene serine/threonine-protein kinase (BRAF1 HUMAN) has status "Disease" because of association with cardiofaciocutaneous syndrome [MIM:115150]

Conclusion: The knowledge about possible genetically determined polymorphisms in protein sequences can be used in iterative database searching of MS data to improve the quality of protein identification. The proposed knowledge-based iterative approach applied for the analysis of the PRIDE MS-data demonstrated its efficiency in revelation of the SAPs in human proteins.

Availability: available from the authors on request.

Acknowledgements: This work was supported by "Proteomics in medicine and biotechnology" program.

USING THE COMPUTER-BASED IMAGE PROCESSING TECHNIQUE IN GENETIC ANALYSIS OF LEAF HAIRINESS IN WHEAT TRITICUM AESTIVUM L.

A.V. Doroshkov*[1], M.A. Genaev[1], T.A. Pshenichnikova[1], D.A. Afonnikov[1,2]
[1]- *Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia*
[2]- *Novosibirsk State University, Novosibirsk, Russia*
e-mail: ad@bionet.nsc.ru
[*]*Corresponding author*

Key words: *wheat, leaf hairiness, trichome*

Motivation and Aim: Leaf hairiness in wheat is of great importance for protection from pests and for adaptation to environmental factors. For example, this trait is characteristic of a number of drought resistant wheat cultivars referred to the steppe ecological group. Study of the features of leaf hairiness morphology and identification the corresponding genes will allow to obtain varieties resistant to hard climatic conditions and certain pests. To identify the genes responsible for the leaf hairiness, mass analysis of a great number of plants belonging to different hybrid populations is needed, accompanying with a laborious manual job.

Methods and Algorithms: Furthermore, a more accurate description of the morphological properties of the trait for correct determination of phenotypic classes is timely. Using of new computer–based technologies for descriptions of quantitative characteristics of leaf hairiness is the important step in this direction. In the course of the work, we used the LHDetect program for determining the degree of leaf hairiness and its morphological properties on the basis of its microscope image processing. The suggested method appeared to be the effective approach for a large scale analysis of leaf hairiness morphological peculiarities in individual plants. For example in according with genotyping this approach can be useful to quantitative trait loci (QTL) mapping.

Results: In this study we detailed analysis of hairines in wheat as a complex feature. For two different cultivars with similar leaf hairiness was shown differences. The disjoining of hairness trait in F2 generation hybrids was studied for several combinations of parents. This allowed us to qualitatively estimated the possible number of genes that may control the hairness trait in different cultivars. It was shown that this trait for several cultivars is polygenic. We show correspondence between trichome middle length and number of trichomes on several cultivares.

Conclusion: LHDetect is fast and powerful tool for analysing leaf hairiness in wheat.

Availability: The LHDetect system is available at http://www.wheatdb.org/lhdetect

The work was supported by RAS Program 2 «Origin and evolution of biosphere»

METAGENOMICS ANALYSIS OF THE BACTERIAL COMMUNITY ASSOCIATED WITH DIATOM ALGA *SYNEDRA ACUS*

Yu. P. Galachyants
Limnological Institute SB RAS, 3 Ulan-Batorskaya Street, Irkutsk, 664033 Russia
e-mail: yuri.galachyants@lin.irk.ru

Key words: *metagenomics, diatoms, pyrosequencing*

Motivation and aim: Metagenomics analysis now becomes a powerful and widely used tool to decipher complex bacterial communities from specific microbiomes and environmental samples. Metagenomics approach gets new breadth with the advent of HTS technologies that allow to sample many thousands of short DNA fragments from a specimen of the question for reasonable time. In the present report we accessed the taxonomy composition and metabolic properties of the microbial community associated with the unialgal culture of a freshwater araphid diatom *Synedra acus* using 454 pyrosequencing platform.

Methods and algorithms: Two total DNA samples from *S. acus* unialgal culture were analyzed using standard shotgun library preparation with Titanium chemistry followed by GS FLX sequencing (Centre "Bioengineering", RAS, Moscow). First DNA sample (CM1) was isolated from untreated *S. acus* culture. Second DNA sample (CM2) was isolated from the *S. acus* culture filtered through 5 μm polycarbonate filter. During this step, a considerable amount of bacteria had been removed that were not tightly bound to *S. acus* cells. Taxonomic classification of metagenomic reads was performed by STAP pipeline using 16S rRNA sequences. The STAP output was processed by in-house developed scripts to generate final taxonomic dendrogams. Metagenomic reads were compared with proteins from the "non-redundant" database using BLASTX program. Resulted BLAST output files were analyzed in MEGAN package to map functional annotation GO terms.

Results and conclusion: As revealed by STAP taxonomic classification of rRNA sequences, the microbial community contains bacteria from wide range of taxa including phylotypes Bacteroidetes, α-, β-, and γ-Proteobacteria. Noticeably, the taxonomic composition of CM1 substantially differs from that of CM2. The most dramatic change is 30-fold decrease of γ-Proteobacteria in CM2. Two other notable points of difference between metagenomic samples CM1 and CM2 are the diversity within α-Proteobacteria and Bacteroidetes groups. Patterns of taxonomic composition in CM1 and CM2 moderately support a hypothesis that α-Proteobacteria utilize cell-bound polysaccharides whereas β-Proteobacteria feed on dissolved polymers and degrade dead algal cells.

We performed an initial analysis of CM2 sample to screen the bacterial enzymes that degrade complex polysaccharides produced by *S. acus*. The CM2 sample was found to contain genes participating in catabolism of plant-like polysaccharides such as pectin, alginic acid, cellulose, chitin, and xylan. This data strongly suggests that there is a part of microbial community that is adapted to use the *S. acus* carbohydrates as the primary source of the organic carbon.

Acknowledgements: This work was financially supported by RFFI grant #09-04-12231 ofi-m. We thank distributed computing group of IDSTU SB RAS and particularly Ivan Sidorov in help to perform CPU-consuming stages of the analysis.

BioinfoWF — WEB SERVICES AND WORKFLOW MANAGEMENT FOR BIOINFORMATICS ANALYSIS

M.A. Genaev*[1], D.A. Afonnikov[1,2]
[1] *Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia*
[2] *Novosibirsk State University, Novosibirsk, Russia*
e-mail: mag@bionet.nsc.ru
[*]*Corresponding author*

Key words: *bioinformatics, workflow, grid processing, XML*

Motivation and Aim: The analysis of biological data in bioinformatics usually consists of several steps performed by different programs subsequently. During the analysis progress, the output of one calculation module serves as an input of the other module, etc. Thus, the overall procedure could be organized as a workflow [1, 2]. For example, the calculation of the phylogenetic tree for protein family requires protein sequence extraction from databases, multiple sequence alignment, phylogeny estimation. It should be noted, that most of single steps could be performed using different routines. For example, sequence alignment could be obtained using ClustalW, Mafft, Muscle or T-Coffee programs. The program's choice by user often depends on the data under analysis and the aim of the task.

Methods and Algorithms: To perform workflow data processing for bioinformatics we developed BioinfoWF system. It is written in Perl and based on the XML description of the program options, input and output data for a single step of the workflow. The second part of the system describes the workflow scheme, set the file data, the execution status of each step. The BioinfoWF runs under command line on the UNIX-like systems or as a web-service. The workflow or its part can also perform on the multiprocessor cluster systems under Sun Grid Engine.

Results: We used BioinfoWF to develop Computer System for Analysis of Molecular Evolution Modes of Protein Families (SAMEM) and functionally important SNP detection in the regulatory regions of eukaryotic genes.

Conclusion: The BioinfoWF can be used to organize workflow management for various bioinformatics tasks.

Availability: The BioinfoWF available at *http://pipeline.bionet.nsc.ru*.

Acknowledgements: This work supported by SB RAS Integration project 1, 26, 113, 119, RAS programs №22 (project 8) and "Biosphere origin and evolution".

References:

1. Rhos J., Karlsson J. and Trelles O. (2009) Magallanes: a web services discovery and automatic workflow composition tool, *BMC Bioinformatics*, **10**:334.
2. Oinn T., Addis M., Ferris J., Marvin D., Senger M., Greenwood M., Carver T., Glover K., Pocock M.R., Wipat A., Li P. (2004) Taverna: a tool for the composition and enactment of bioinformatics workflows, *Bioinformatics*, **20**:3045-3054.

INTEGRAL EVALUATION OF PROFILE OF NATRIURETIC PEPTIDES SYSTEM'S MRNA EXPRESSION IN CULTURED CARDIOMYOCYTES DURING ANOXIA-REOXYGENATION

V.L. Gurianova*, V.E. Dosenko, V.B. Pavlyuchenko, A.A. Moybenko
Bogomoletz Institute of Physiology, Kiev, Ukraine
e-mail: nika.biph@gmail.com

Motivation and Aim: Natriuretic peptides system (NPS) is a powerful cardioprotective mechanism in mammalians and consists of atrial (ANP), brain, or type B (BNP) and type C (CNP) NPs, and their receptors: NPR-A (for ANP and BNP), NPR-B (for CNP) and NPR-C (clearance receptor eliminating all types of NPs and generating additional signal). This system is in the limelight from the moment of its finding and a lot of studies are carried out in the field of genetic mechanisms underlying the NPs' regulation and their role in cardiac physiology and pathology but more data bear much more questions because of their contradictoriness and uncomplexity. This fact caused our interest in integral evaluation of members of NPS expressed by cardiomycytes on genetic level in different conditions in cultured cardiomyocytes including such important pathological process as anoxia-reoxygenation (AR) as a model of ischemia-reperfusion.

Methods and Algorithms: The study was carried out on neonatal cardiomyocytes isolated from ventricular myocardium of 2-day-old Wistar rats. Anoxia (A) was attained with an airtight jar from which the O_2 was flushed with gas mixture containing 5% CO2 and 95% Ar for 30 min. Reoxygenation (R) was realized by exchanging fresh medium and by its aeration with the standard gas mixture for 60 min (acute AR) or 24 hours (prolonged AR). Total RNA was isolated from cells (Trizol RNA-prep kit) with the followed by reverse transcription and quantitative Real-Time PCR (SYBR Green PCR or TaqMan Gene Expression) performing to determine the level of mRNA expression of NPPA (ANP), NPPB (BNP), NPR1 (NPR-A) and NPR3 (NPR-C) genes in control (incubation for 48 hours in normoxic conditions), acute AR and prolonged AR groups. Except the valuables of obtained results in themselves we tried to find the most optimal coefficient to estimate the NPS's general activity in every individual case and in dynamics of changes. So, we assume that all changes of mRNA level are realized in adequate peptides and proteins changes, the probability of ligand-to-receptor binding is in direct proportion to ligand-to-receptor affinity (or in inverse proportion to dissociation constant (B.Bennett, 1991)) and that the force of effect depends on the level of ANP, BNP and the index NPR1/NPR3 that points at the state of receptor apparatus and NPR3 number also points on the state of clearance potential of the system and is denominator of coefficient. So, NPS coefficient = S Ч NPPA Ч ((0.53 Ч NPR1) / (0.38 Ч NPR3)) + S Ч NPPB Ч ((0.14ЧNPR1) / (0.08 Ч NPR3)).

Results: Received data show that NPPA, NPPB and NPR1 mRNA expression in all studied by us cases changed in the same manner: extremely augmented during the acute AR and backing to the control values (and even lower) during the prolonged R. As opposed to this, NPR3 gene mRNA expression changed in the opposite direction that forms the additive contribution to general tendency of changes of NPS's general activity that we express in form of relevant coefficient (in control group it is equal 0.79 ± 0.2, after acute AR it was increased to 16.03 ± 7.89 ($P < 0.05$, compared to control) and after prolonged AR came back to control values and not significantly lower by them – 0.41 ± 0.14 ($P > 0.05$, compared to control).

Conclusion: We obtained the data about the changes of mRNA expression of members of natriuretic peptides system expressed by cardiomyocytes in conditions of acute and prolonged AR compared to control and are able to conclude that these changes have similar manner for natriuretic peptides end effector receptor and opposite reaction for clearance receptor that is additive for the change (increased or decreased) of NPS's general activity on genetic level.

VARIABLE PART OF GENE EXPRESSION PROFILES IN LIVER AND KIDNEY OF PIGS

N.S. Khlopova*, T.T. Glazko
Russian State Agrarian University – MTAA named after K.A.Timiryazev, Moscow, Russia
e-mail: hns86@mail.ru
**Corresponding author*

Key words: *microarrays, gene expression profiles, key genes, variable expression.*

Motivation and Aim: Estimation of organ-specific gene expression profiles using DNA microarrays is used for searching the genes which transcription can bring in the key contribution to organ metabolome. Usually such analysis excludes the genes with individual differences in expression between the investigated animals. In order to find out the possible reasons of such variability we carried out the analysis of the interrelationships between genes, expression of which showed the essential differences in transcription level in a liver and kidney between six Landrace pigs.

Methods and Algorithms: The analysis of hybridization intensity of cDNA, synthesized on the mRNA of the liver and kidney of six Landrace pigs was carried out on DNA microarrays consisting of 19980 of 70-mer oligonucleotide probes [1]. The arrays were scanned on GSI Lumonics ScanArray 5000 laser scanner. Mathematical processing was carried out with Statistica.

Results: The analysis of 600 genes, distinguished by hybridization intensity between liver and kidney on microarray probes for more than on 20000 standard units of the luminescence in investigated pigs allowed forming two groups of the genes. The first group included genes with the same contribution in gene expression profiles in different animals ("constitutive" genes) and the second group joined the genes with varied expression levels in different animals ("variable" genes). A total of 24 genes with the significant individual variability in expression were revealed. The individual animal variability in gene expression correlated between liver and kidney in 11 of these 24 genes ($r > 0,96$; $P < 0,05$). That was allowed to assume the presence of common regulatory factors for both organs for these 11 genes. Subdividing of 24 variable genes into groups with inner statistically significant correlations in individual variability of gene expression was revealed. As a rule, genes in individual groups with the interconnected expression belonged to the general metabolic way. Totally 8 groups were allocated - 6 genes which products participated in formation of a blood clot; 6 genes - in transport and a lipid metabolism; 5 genes represented markers of blood cells and lysosomes; products of 3 genes participated in programmed cell death and four groups in 2 genes which products participated in Ca^{2+} transport, in intercellular matrix creation, reflected the mitochondrion functional activity and the hormone depended genes.

Conclusion: The obtained data testified that individual variability in gene expression between animals could be caused by the variability of regulatory factors external for organs and different blood fullness of organ samples.

References:

1. Glazko T.T., Khlopova N.S., Fahrenkrug S., Glazko V.I. (2009) Gene expression profiles in liver and kidney of pigs, Izvestia of Timiryazev Academy, Special Issue: p. 55-60

DWARF FORM OF *MALUS BACCATA* (L.) BORKH: INITIAL STAGE OF THE PARAPATRIC SPECIATION?

E.V. Kuznetsova*[1], T.E. Peretolchina[2], A.V. Rudikovskiy[1], D.Yu. Sherbakov[2,3]
[1]*Siberian Institute of Physiology and Biochemistry of Plants SB RAS, Irkutsk, Russia*
[2]*Limnological Institute SB RAS, Irkutsk, Russia*
[3]*Irkutsk State University, Faculty of Biology and Soil Science*
e-mail: elenakuznetsova01@gmail.com
**Corresponding author*

Key words: *Apple-trees, M. baccata, dwarf forms, microsatellites, speciation.*

Motivation and Aim: Many authors consider the parapatric speciation as important way of speciation among animals, however, it is not so common in plants. One rare example of the possible parapatric speciation in plants is Siberian apple tree *Malus baccata* (L.) Borkh and its rare form dwarf apple tree which is found on the territory of the Republic of Buryatia in the contact zone of forest and steppe. Taxonomic status of the dwarf apple tree remains undecided. So the aim of our investigation was determining of the phylogenetic relationships between different forms of Siberian apple trees and clarifying of the origin of the dwarf apple tree.

Methods and Algorithms: The samples were collected on the territory of the Republic of Buryatia. We chose four geographic locations of plants: №1 is mixed, it consists of dwarf and tall forms of the apple-trees and grows in a valley of Zagustaj river in Republic Buryatia. The second group is represented by the dwarf forms growing near Yagodnoe village near Gusinoozersk along a stream. The group №3 is natural typical tall forms of *M. baccata* that occupy the territory around Kabansk village. Group №4 is man-made planting of tall forms in the Yagodnoe village. Genomic DNA was extracted by the modified methods of Doyle and Dicson from leaves.

Six pairs of specific micosatellite primers were used for analysis of genetic diversity within the studied groups of plants: 01ab, 02b1, 04H11, 05G8, 23G4, 28F4. Genetic structure of studied apples-trees was determined using software STRUCTURE 2.3.1. (number of populations K from 2 to 5, burnin period – 50 000, number of MCMC reps after burn – 1 000 000). Phylogenetic tree based on micosatellite loci was reconstructed using distance matrix DAS.

Results: Microsatellite phylogenetic tree has shown that the all studied groups of dwarf and tall forms of Siberian apple tree form independent clades and also dwarf forms is separated from tall plants. Dwarf forms look polyphyletic on the tree.

Conclusion: At the edge of the areal the differences between the specific habitats are quite high and may play a role in a selection, which is reflected in the genetic differentiation of various growth forms of apples. On the other hand, only extreme biotopes (for example, highly arid places) where there is a dwarf form, can favor to the survival of such marginal forms which, apparently, are more adapted for these conditions. The question of the heritability of dwarfism remains open, due to the long duration of generations within apple trees. The phylogenetic analysis has shown that the dwarf form has descended from tall *M. baccata*. This fact suggests that dwarf apples are ecological forms of the Siberian apple which, probably, represents the initial stage of the parapatric speciation.

COMPUTER SYSTEM SITEX FOR ANALYZING PROTEIN FUNCTIONAL SITES IN EUKARYOTIC GENE STRUCTURE

I.V. Medvedeva*, P.S. Demenkov, V.A. Ivanisenko
Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
e-mail: brukaro@bionet.nsc.ru
**Corresponding author*

Key words: *protein functional sites, database, eukaryotic gene structure*

Motivation and Aim: Analyzing protein structure projection on exon-intron structure of corresponding gene through years led to several fundamental conclusions about structural and functional organization of the protein [1]. According to these results we decided to map the protein functional sites and analyze the specificity of the coding positions. Although there are some databases (SEDB, ExDom, XdomView) that store the information about protein structure mapping, there is no information about protein functional sites with one-to-one scheme of projection.

Methods and Algorithms: We integrated the resources from PDB and Ensembl to match protein tertiary structure and corresponding gene as one-to-one projection. Also we added the descriptive information about domains (from Pfam) and protein classification (from SCOP) and map their borders. To evaluate the protein functional site discontinuity we counted two coefficients for each of them: one coefficient measures the discontinuity in protein sequence and the other between exons, encoding the site. So we collected for about 10 000 unique sites from 2 500 unique sequences from 17 organisms. Among this we included the BLAST search and 3D similar structure search using PDB3DScan for the polypeptide encoded by one exon, participating in organizing the functional site.

Results: Analyzing protein functional site distribution through exon structure we found that the aminoacid usage is not the same along the length of the exon. Thus, the most frequent codon nearby exon-intron boundaries is TAT (encode tyrosine). Also it was found the usage of non-optimal codon for some aminoacids for 5`- and 3`-ends. The statistical analysis also revealed that protein functional sites are mostly encoded by the first exons of the sequence.

Conclusion: We created the computer system SitEx to help: 1) to study the positions of the functional sites in exon structure; 2) to make the complex analysis of the protein function; 3) to exposure the exons that took part in exon shuffling and came from bacterial genomes; 4) to study the peculiarities of coding the polypeptide structures.

Availability: http://samurai.bionet.nsc.ru/~demps/sitex/

Acknowledgements: This work was supported by State contract FASI №02.514.11.4123, Interdisciplinary integration project SB RAS №111, №94, Program RAS 22. Molecular and cellular biology.

References:

1. H. Kaessmann, S. Zöllner, A. Nekrutenko, W. H. Li. (2002) Signatures of domain shuffling in the human genome. Genome Res. 2002. 12: 1642-1650.
2. Perrière G., Thioulouse J. (2002) Use and misuse of correspondence analysis in codon usage studies. Nucleic Acids Res. 30:4548–4555.
3. Abelson J., Trotta C.R., Li H. (1998) tRNA splicing. J Biol Chem 273:12685–12688.

PRION-ASSOCIATED PROTEINS IN YEAST: COMPARATIVE ANALYSIS OF YEAST STRAINS, DISTINGUISHED BY THEIR PRION CONTENT.

E.V. Mikhailova*, O.V. Nevzglyadova, A.V. Artyomov, T. R. Soidla
Institute of Cytology RAS, Saint-Petersburg, Russia
e-mail: mikhailovaev@list.ru
**Corresponding author*

Key words: *prion, amyloid, heterokaryon, cytoduction, Saccharomyces cerevisiae*

Motivation and Aim: systematic detection and identification of yeast proteins, that can associate with prions, forming the so-called prion aggregates.

Methods and Algorithms: our approach to the identification of prion-associated proteins includes isolation of the pellet fraction, enriched by prion aggregates of yeast cells crude lysates and comparison of isogenic yeast strains, differing only by their prion composition. These strains were obtained using the cytoduction and various prionotropic treatments. To get [*PSI*$^+$] strains we applied overexpression of the *SUP35* gene located on a plasmid. To obtain [*psi*$^-$] strains we used GuHCl treatment or overexpression of the plasmid-born *HSP104* gene. 2D electrophoresis followed by MALDI analysis of the pellet proteins of [*PSI*$^+$] and *[psi*$^-$] strains permitted identification of the prion-associated proteins.

Results: More than 30 proteins whose presence in the pellet fraction correlate with the change of prion(s) content were identified. Approximately a half of these proteins belong to chaperons and to enzymes of glucose metabolism. Chaperons are known to be involved in prion metabolism and are expected to be present in prion-containing aggregates. Nevertheless, several recent data suggest that the presence of glucose metabolism enzymes is not accidental too (3). We also detected six proteins involved in oxidative stress response, eight – in translation, and several proteins involved in proteolytic degradation.

Conclusion: We would like to conclude that our method seems to reveal the same core prion-associated proteins, as other approaches (1, 2, 3). At the same time our approach is less restrictive and has enabled us to identify some additional proteins – all of them likely to be a respond to signals of protein damage and misfolding. Most of identified proteins seem to be prion-associated, but we can't exclude the possibility that several proteins may propagate as prions.

References:

1. Bagriantsev S.N., Gracheva E.O., Richmond J.E., Liebman S.W. 2008. Variant-specific [*PSI*$^+$] infection is transmitted by Sup35 polymers within [*PSI*$^+$] aggregates with heterogeneous protein composition. Mol. Biol. Cell. 19 : 2433-2443
2. Wang Y., Meriin A.B., Costello C.E., Sherman M.Y. (2007) Characterization of proteins associated with polyglutamine aggregates: A novel approach towards isolation of aggregates from protein conformation disorders. Prion, 1 : 128–135.
3. Wang Y., Meriin A.B., Zaarur N., Romanova N.V., Chernoff Y.O., Costello C.E., Sherman M.Y. (2008) Abnormal proteins can form aggresome in yeast: aggresome-targeting signals and components of the machinery. FASEB J., 23 : 451-463.

INFLUENCE OF HYDROPHOBICITY OF BETA-BLOCKERS ON THEIR BINDING AFFINITY

V.N. Novoseletsky*, R.G. Efremov
Institute of Bioorganic Chemistry RAS, Moscow, Russia
e-mail: valeryns@gmail.com
**Corresponding author*

Key words: *beta-blockers, hydrophobicity, binding affinity*

Motivation and Aim: G-protein-coupled receptors (GPCR) play a key role in signal transduction in living cells, participate in regulation of numerous biological functions and, thus, are important therapeutic targets in a variety of disease states. Beta-adrenoreceptor antagonists are among the most widely used drugs in clinical practice. It has been shown earlier that antagonist binding is accompanied by a hydrophobic interaction between receptor and ligand [1]. To further investigate this issue and arrive to numerical models, we used theoretical modeling approach to the estimation of hydrophobic interactions between the beta2-adrenoreceptor and a set of its antagonists with experimentally measured binding constants.

Methods and Algorithms: To calculate hydrophobic interactions we used web-server PLATINUM (http://model.nmr.ru/platinum) [2]. It is an easy-to-use and customizable tool for estimation of the hydrophobic/hydrophilic match or mismatch at the interface of two molecules based on the concept of the Molecular Hydrophobicity Potential (MHP). It utilizes empirical atomic hydrophobicity constants derived from the water-octanol partition coefficients for organic compounds. Recently determined spatial structure of adrenoreceptor [3] was used to perform molecular docking of antagonist molecules. Resulting complexes were then optimized to satisfy experimental spatial restraints. Original approach was proposed for estimation of hydrophobic match of beta-blockers in the binding site of beta2-adrenergic receptor.

Results: Detailed analysis of hydrophobic match/mismatch of the beta-blockers reveals some properties of substituents which are crucial for high binding affinity. Hydrophobicity of particular fragments is used for construction of scoring function for binding affinity.

Conclusion: Strong correlation (R^2 = 0.8) between pKd and hydrophobic match was found. Therefore our method of estimating hydrophobic match can be further used to discover new potent inhibitors to beta2-adrenoreceptor.

Acknowledgements This work was supported by the grant of Russian Foundation for Basic Research № 07-04-01514-a, by the Program RAS MCB and by the grant SS-4728.2006.

References:

1. M.L.Contreras et al. (1986) Thermodynamic properties of agonist interactions with the beta adrenergic receptor-coupled adenylate cyclase system. II. Agonist binding to soluble beta adrenergic receptors, *J Pharmacol Exp Ther,* **237:** 165-172.
2. T.V.Pyrkov et al. (2009) PLATINUM: a web tool for analysis of hydrophobic/ hydrophilic organization of biomolecular complexes, *Bioinformatics,* **25:** 1201-1202.
3. V.Cherezov et al. (2007) High-resolution crystal structure of an engineered human beta2-adrenergic G protein-coupled receptor, *Science*, **318:** 1258-1265.

COMPARISON OF METHODS FOR RECONSTRUCTION OF MODELS FOR GENE EXPRESSION REGULATION

A.A. Shadrin[1,*], I.N. Kiselev[1], F.A. Kolpakov[2,1]
1Technological Institute of Digital Techniques SB RAS, Novosibirsk, Russia
2 Institute of Systems Biology, Novosibirsk, Russia
e-mail: inter.cm@mail.ru
**Corresponding author*

Key words: *gene expression regulation, microarray experiment, experimental data smoothing, parameters optimization.*

Motivation and Aim: DNA microarray technology is widely used nowadays to gain data on gene expression; however, information obtained in this way can not be used directly and needs processing. It would facilitate the simulation of gene interactions and the reconstruction of the process under study, as well as give the opportunity to predict the development of the process at various changes in the conditions of its occurrence. The objectives of this study were: 1) consideration of linear and nonlinear [1] differential models of gene expression regulation, 2) study of various data processing methods on the models behavior, and 3) test and comparison of the considered models on the base of data published in [2].

Methods and Algorithms: For experimental data smoothing we used two different types of smoothing spline, nuclear smoothing with Epanechnikov core and the classic least squares smoothing method. During models parameters optimization we compared three methods: simulated annealing, evolutionary algorithm [3] and advanced method of gra-dient descent.

Results: Two differential models of gene expression regulation were examined in details on the dataset containing yeast cell cycle associated genes expression, measured as amounts of mRNA using microarrays at 18 time points over two cell cycle periods, which was published in [2]. The effect of data smoothing method and parameters optimization algorithm were investigated. The best result for the data smoothing was obtained using a spline with a fixed weighting parameter, in the parameters optimization the best in accuracy and speed proved to be the evolutionary algorithm.

Conclusion: Although the study performed on the example of two models demonstrated that the spline with a fixed weighting parameter and the evolutionary algorithm were the most optimal, it also became apparent, that the choice of methods for data smoothing and parameters optimization could strongly influence the behavior of the model under other conditions, hence, each study requires individual approach to select the most (or more) optimal methods.

Acknowledgements: This work was supported by EU grant No.037590 “Net2Drug”.

References:

1. T.T.Vu, J.Vohradsky. (2007) Nonlinear differential equation model for quantification of transcriptional regulation applied to microarray data of Saccharomyces cerevisiae, *Nucleic Acids Research*, 35: No. 1. 279-287.
2. P.T.Spellman, G.Sherlock, M.Zhang, V.Iyer, K.Anders, M.Eisen, P.Brown, D.Botstein, B.Futcher. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization, *Mol. Biol. Cell*, 9: 3273–3297.
3. T.P.Runarsson, X.Yao. (2000) Stochastic Ranking for Constrained Evolutionary Op-timization, *IEEE Transactions on evolutionary computation*, 4: No. 3. 284-294.

APPLICATION OF DNA SEQUENCING AND DETECTION OF LEVELS EXPRESSION EPIDERMAL GROWTH FACTOR RECEPTOR (EGFR) IN CANCER PRACTICE

S.M. Sirotkina[1], A.H. Sabirov[2]*
Tyumen State University, Tyumen, Russia[1]
Tyumen Academy of Medicine, Tyumen, Russia[2]
e-mail: sabirov58@mail.ru
**Corresponding author*

Key words: *epidermal growth factor receptor, DNA sequencing.*

Motivation and Aim: The aim of the research is to study prognostic and diagnostics significance of the detection expression levels of EGFR in serum, as well as the presence of mutations in certain oncogenes of cancer patients.

Methods and Algorithms: Research is based on testing of cancer patients who were having the examination and therapy in Sabirov Center of molecular and genetic diagnostics from 2008 to 2009. The research includes 220 patients at the ages of 25 to 70 with various malignant tumors: mammary gland cancer (60 patients), stomach cancer (75 patients), lung cancer (25 patients) and colon cancer (60 patients). The level of loose EGFR forms in serum was tested by means of immunoferment analysis (IFA) in ELISA modification.

Material for molecular genetic studies provided the tumor cells and blood plasma of these patients. The presence of mutations was determined in the oncogenes p53 (exons 5,6,7,8), C-kit, B-raf, APC, K-ras, E-cadherin and p16 by means specially chosen primers. The results verified by direct sequencing of PCR fragments.

Results: The average value of level of loose EGFR forms in serum of patients with mammary gland cancer was 4,47 fmol/ml. For patients with stomach cancer this level was 4,8 fmol/ml, for lung cancer patients it was 4,77 fmol/ml (an average standard of EGFR is 0-3,6 fmol/ml).

The average value of level of loose EGFR forms in serum of patients with colon cancer was 19,6 ng/ml (an average standard of EGFR is 0-9,6 ng/ml).

When molecular genetic studies have observed a significant decrease in the number of mutations in the course of special treatment after 2 courses of chemotherapy (PCT) (7 cases). After 6 courses of PCT, mutations persisted in 4 patients, i.e. molecular remission was not achieved, indicating a lack of elimination of tumor cells, unreasonable transfer of the patient in 3 clinical group and required PCT need to continue or change the scheme.

Conclusion: 1.There were excess of upper bounds of loose EGFR forms expression level standart (overexpression) in the course of research.

2. So it is reliable that EGFR overexpression is of importance in carcinogenesis. It is a marker for describing tumor behavior and instituting individual treatment for patient with malignant tumors.

3. EGFR overexpression, as well as the presence of mutations in certain oncogenes of cancer patients (p53 (exons 5,6,7,8), C-kit, B-raf, APC, K-ras, E-cadherin and p16) can also be a target for creation of new kinds of anti-tumor therapy pointed at mitosis signal transmission blocking (for targeting therapy).

SEQUENCE ANALYSIS OF COG3868 and COG2342 FAMILIES

O.O. Stepuschenko[1], D.G. Naumoff*[2]
[1] *Kazan State University, Department of Genetics, Kazan, Russia*
[2] *Institute for Genetics and Selection of Industrial Microorganisms, Moscow, Russia*
e-mail: daniil_naumoff@yahoo.com
[*] *Corresponding author*

Motivation and Aim: The endo-α-1,4-polygalactosaminidase (EC 3.2.1.109) is a very rare enzyme, which has been found only in two strains of bacteria (Streptomyces griseus C-10 and Pseudomonas sp. 881). It belongs to the GH114 family of glycoside hydrolases (or COG3868). According to the CAZy database (http://www.cazy.org/), this family includes 56 proteins. COG3868 is closely related to functionally uncharacterized COG2342. The latter includes proteins with TIM-barrel type of the three-dimensional structure (PDB, 2AAM). This type of 3D structure has been shown for the catalytic domains of four clans (a hierarchical group higher than family) of glycoside hydrolases. Relationships within the GH114 family are still unclear and became the purpose of the work, as well as its evolutionary connections with other protein families.

Methods and Algorithms: Protein sequences were retrieved from the NCBI database. Multiple sequence alignment of 117 GH114 domains was made in BioEdit program (very similar and partial sequences were omitted). The phylogenetic trees were built using programs of PHYLIP package. 34 members of COG2342 family were used as outgroup. Interfamily relationships were established using PSI Protein Classifier. This program analyzes results of PSI-BLAST searches. Several most divergent representatives from the GH114 and COG2342 families were used as a query.

Results and Conclusion: We have revealed 130 non-identical protein sequences of GH114 domains using the blast algorithm. They include representatives of several phyla of Bacteria (Actinobacteria, Aquificae, Chloroflexi, Deferribacteres, Deinococcus, and Proteobacteria), as well as some Eukaryota (Alveolata, Ascomycota, Basidiomycota, Chlorophyta, and Oomycetes). The majority of the proteins have similar length (251–375 amino acid residues) and contain only one domain (GH114). Protein from *Hahella chejuensis* (GenPept, ABC28688.1) consists of two GH114-domains; protein from *Endoriftia persephone* (ZP_02533448.1), in addition to GH114, has a GH9-domain.

Three main clusters can be recognized on the GH114 phylogenetic trees (both Neighbor-Joining and Maximum Parsimony). Two of them have high bootstrap support and include mainly representatives of Actinobacteria. The third cluster is formed by proteins from Ascomycota, but they are not very well separated from some bacterial proteins. The tree topology supports the important role of horizontal transfer in the evolution of GH114 proteins. Two very conserved residues (Asp and Glu) most probably are the nucleophile and proton donor in the active center, respectively.

Iterative screening of the protein database allowed us to reveal relationship of GH114 and COG2342 with GH5 (clan GH-A), GH13 (clan GH-H), GH18 (clan GH-K), GH20 (clan GH-K), GH27 (clan GH-D), GH29, GH31 (clan GH-D), GH35 (clan GH-A), GH36 (clan GH-D), GH42 (clan GH-A), GH66, GH97, GH101, COG1306, COG1649, GHL3, and GHL4 families. These data support the common evolution origin of all TIM-barrel type glycoside hydrolase catalytic domains, as we suggested earlier (D.G. Naumoff, BGRS'2006).

THE CORRESPONDENCE BETWEEN 5'-UNTRANSLATED REGIONS OF YEASTS'S GENES WITH THEIR EXPRESSION EFFICIENCY.

A. Orlenko
Novosibirsk State University, Novosibirsk, Russia
e-mail: desmidium@yandex.ru

Key words: *efficiency elongation index, gene expression, nucleosom potential*

Motivation and Aim: studying of gene expression efficiency is actual and essential problem for theoretical and practical application. There are a lot of factors which make an influence on expression rate on translation level: gene codon composition, the presence and the distribution of second structure in mRNA, "the strength" of those structures. Gene expression efficiency is also depends on localization of nucleosom in 5'- untranslated regions. The goal of the present research consist in the finding of correspondence between elongation efficiency index (EEI) and nucleosom potential in 5'-untranslated region in two yeast's genome: *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*.

Methods and Algorithms: Sequences from -600 to +600 nucleotide relative to ATG codon of yeast's genes were derived from gene cards of yeasts organisms, which were taken from data base GeneBank. Then sequences were sorted in accordance with EEI in descending order. EEI was calculated with using EEI-Calculator program. On the second stage a nucleosom potential was computed for sorted sequences with using RECON program. For each sequence profile which was consisted from 1061 nucleosom potential values was formed. Then a correlation coefficient between nucleosom potential values in each of 1061 positions of all gene set and EEI of corresponding genes was found. For obtained results both a Student's and a Fisher's test significance was calculated for such probability values as 0.95, 0.99, 0.999.

Results: a significant negative correlation between nucleosom formation and EEI with 0.95 probability was found: for all 5649 genes of *Saccharomyces cerevisiae* in following regions (-230;-345), (-450;-480) and (-510;-600), for 10% of *Saccharomyces cerevisiae* genes with the highest EEI values for some positions in (-230;-340) interval, for all 4546 genes of *Schizosaccharomyces pombe* for the whole sequence region, for 10% of all *Schizosaccharomyces pombe* genes with the highest EEI values in (-270;-460) and (-95;461) region. For 10% of all *Schizosaccharomyces pombe* genes with the lowest EEI value a significant positive correlation was found almost at a whole sequence length. For experimentally verified 231 *Saccharomyces cerevisiae* genes it was found that with 0.95 probability for approximately half of all numbers of positions in (-194;-171) region a significant negative correlation between EEI and nucleosom potential exists.

Conclusions: results allowed making several suggestions: for *Saccharomyces cerevisiae* genes with high and moderate expression level several regions with the low probability of nucleosom formation, including one near translation start site, promote expression, for *Schizosaccharomyces pombe* genes with high and moderate expression level 5'-untranslated region with a low probability of nucleosom formation also promote expression. Also it is possible to suggest that for genes with the lowest EEI values the presence of nucleosom in 5'-untranslated region and in the region of TSS blocks expression. For an experimentally obtained promoter sequences is also possible to suggest that for effective expression of those genes the presence of sites with low probability of nucleosom formation is necessary.

GPGPU-COMPUTING FOR PREDICTION OF SMALL LIGAND BINDING SITES IN PROTEINS

L.A. Uroshlev[1], I.V. Kulakovskiy [2], S.V. Rakhmanov[2], V.M. Makeev[2]
[1]*Moscow State Forest University, Mytischi, Russia*
[2]*Research Institute for Genetics and Selection of Industrial Microorganisms, Moscow, Russia*
e-mail: leoniduroshlev@gmail.com

Key words: *empirical potentials, ion binding, GPGPU-computing*

Motivation and Aim: At this time there are over 60 thousand known protein 3D structures and this number is ever increasing. One can use this data to obtain empirical potentials for further prediction of small ligand binding sites in a protein structure of interest [1]. This approach is computationally expensive which often limits its practical usage to small molecular structures.

Methods and Algorithms: Here we present an effective strategy to use GPGPU for prediction of small ligand binding sites in proteins. We took the method from [2], which was successfully applied to predict water- and calcium ions in a number of different protein structures. We developed a parallel version of the algorithm and implemented it using CUDA [3] programming model.

For each protein structure CUDA architecture allows to divide it into hundreds of independent blocks for parallel processing. Taking a list of protein structures containing a ligand of interest we compute all contants between the ligand and all other atoms of the structure and store this data in the simple relational database. Then it can be used to evaluate observed and expected distance frequencies and therefore to produce an empirical potentials for the selected ligand.

Taking an independent protein structure the obtained empirical potentials can be futher used to predict possible ligand binding sites. Here we use CUDA to calculate estimates of local ion binding probability for either randomly placed points (Monte-Carlo simulation mode) or fixed-step grid in a parallel mode.

Results: We present a software PIONCA (Protein-ION CAlculator) which uses effective GPGPU implementation of the empirical potential-based method for small ligand binding prediction. We tested it using data for different metal ions and water.

Conclusion: We present an effective computational tool which shows significant performance gain over initial implementation [1].

Availability: The software is available online at the following URL http://line.imb.ac.ru/ion-calculator.

References:

1. S. Rahmanov, V. Makeev (2007) Atomic hydratation potentials using a Monte Carlo Reference State (MCRS) for protein salvation modeling, *BMC Structure Biology,* **7:19**.
2. S. Rahmanov, I. Kulakovskiy, L. Uroshlev, V. Makeev. (2009) Empirical potential for ion binding in proteins, *Journal of Bioinformatics and Computation Biology*. In press
3. NVIDIA CUDA™ Programming Guide SDK, http://developer.download.nvidia.com/compute/cuda/2_2/toolkit/docs/NVIDIA_CUDA_Programming_Guide_2.2.pdf

Научное издание

Материалы международной школы молодых ученых «Биоинформатика и системная биология» (Новосибирск, 2010)

на английском языке

Proceedings of the young scientists school “Bioinformatics and Systems biology” (Novosibirsk, 2010)

Abstracts have been printed without editing as received from the authors

Школа проводится при поддержке Федерального агентства по науке и инновациям (02.741.12.2181) и Российского фонда фундаментальных исследований (10-04-06022-г)

Подготовлено к печати
в редакционно-издательском отделе
Института цитологии и генетики СО РАН
630090, Новосибирск, пр. акад. М.А. Лаврентьева, 10

Дизайн и компьютерная верстка: А.В. Харкевич

Подписано к печати 17. 06. 2010 г.
Формат бумаги 70 ×108 1/16. Печ. л. 3,1. Уч.-изд. л. 3,1
Тираж 80.

Отпечатано в Институте цитологии и генетики СО РАН
630090, Новосибирск, пр. акад. М.А. Лаврентьева, 10