THE FEDERAL RESEARCH CENTER INSTITUTE OF CYTOLOGY AND GENETICS
SIBERIAN BRANCH OF THE RUSSIAN ACADEMY OF SCIENCES

THE 7th INTERNATIONAL YOUNG SCIENTISTS SCHOOL "SYSTEMS BIOLOGY AND BIOINFORMATICS" SBB-2015

Abstracts

Novosibirsk, Russia
June 22–25, 2015

International Program Committee

Prof. Nikolay A. Kolchanov	Academician RAS, Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia (Chairman)
Prof. Klaus Mayer	Helmholtz Zentrum Muenchen, Neuherberg, Germany
Prof. Inna Lavrik	Otto von Guericke University, Magdeburg, Germany
Prof. Evgenii Rogayev	University of Massachusetts, USA
Dr. Maxim Zakhartsev	Hohenheim University, Institute of Plant Physiology and Biotechnology, Germany
Dr. Yuriy Orlov	Novosibirsk State University, Russia
Prof. Dmitriy Sherbakov	Limnological institute, Irkutsk, Russia
Dr. Irina Abnizova	Wellcome Trust Sanger Institute, UK
Prof. Rene te Boekhorst	University of Hertfordshire, UK
Prof. Ancha Baranova	George Mason University, USA
Prof. Haihua Bai	Inner Mongolia University for the Nationalities, China
Prof. Guoliang Li	Huazhong Agricultural University, Wuhan, China
Prof. Ming Chen	Zhejiang University, Hangzhou, China

Local Organizing Committee

Ilya Akberdin, PhD
Dmitry Afonnikov, PhD
Nadezhda Glebova
Tatyana Karamysheva, PhD
Galina Kiseleva, PhD
Ekaterina Kulakova
Andrey Kharkevich
Sergey Lashin, PhD
Victoria Mironova
Anna Onchukova
Yuriy Orlov, PhD
Olga Popik
Erlan Tokpanov
Svetlana Zubova (chair)

Contacts

The Federal Research Center Institute of Cytology and Genetics
The Siberian Branch of the Russian Academy of Sciences
RUSSIAN FEDERATION
630090, Novosibirsk
Prospekt Lavrentyeva 10
Tel: +7(383) 363-49-80
Fax: +7(383) 333-12-78
E-mail: sbb2015@ig.sbras.ru
URL: http://conf.nsc.ru/SBB2015/ru

The School was financially supported from Russian Science Foundation grant "Programmed cell death induced via death receptors: Delineating molecular mechanisms of apoptosis initiation via molecular modeling" № 14-44-00011.

ISBN 978-5-91291-023-4

CONTENTS

STATE TRANSITION DIAGRAMS IN HYPOTHETICAL MODELS OF CYCLIC GENE NETWORKS FUNCTIONING

A.A. Akinshin[1], M.V. Kazantsev[2]*
[1]*Sobolev Institute of Mathematics SB RAS, Novosibirsk, Russia*
[2]*Altai State Technical University named after I.I. Polzunov, Barnaul, Russia*
[*]*e-mail: markynaz.astu@gmail.com*
[*]*Corresponding author*

Motivation and Aim: Existence and location of periodical trajectories are important properties of multi-dimensional dynamical systems. Suggested approach allows us to identify some important properties of dynamical systems that model cyclic gene networks with positive and negative feedbacks. These properties may be used to describe some important biological properties of these gene networks, such as existence of biorhythms. Acquired results can also be used for analyzing of some autorepressor models.

Methods and Algorithms: We consider non-linear multi-dimensional dynamical systems of chemical kinetics that model hypothetical cyclic gene networks:

$$\dot{x_1} = f_1(x_n) - x_1, \quad \dot{x_i} = f_i(x_{i-1}) - x_i, \quad i \in \{2 \ldots n\}$$

Here x_i represent concentrations of reacting enzymes. f_i are smooth or step functions. Some of them are non-increasing (L-functions) and the rest are non-decreasing (Γ-functions). This describes negative and positive feedbacks in the gene network. A cyclic trajectory in such system corresponds to biorhythm in the modelled gene network (see [1]). Let x_0 be an equilibrium point of the dynamical system. We split its invariant neighborhood into 2^n parallelepipeds by hyperplanes that are parallel to the coordinate hyperplanes and contain the point x_0. These parallelepipeds are called **domains** and they form a graph called **transition state diagram**. A periodic trajectory of the dynamical system is contained in those domains that form a loop in this graph (see [2]).

We define **domain potential** as amount of outcoming edges of this domain in the state transition diagram. We define potential level as union of all domains with the same potential. All domains of a cycle always have the same potential, therefore they belong to one potential level. We prove that all levels except the ones with potential n and 0 contain at least one cycle. This defines the required and sufficient conditions of existence of cycles on potential levels of the transition state diagram.

The main object of our studies of transition state diagrams is parity of amount of Γ-functions in the set of the functions f_i. For each n, all n-dimensional systems can be separated into two equivalence classes. All systems with odd number of Γ-functions belong to one class and the rest belong to the second. State transition diagrams of two systems from the same class are isomorphic.

Results: We obtain necessary and sufficient conditions of existence of cycles in a state transition diagrams, and we prove that they are isomorphic if the parity of the amounts of Γ-functions coincide.

Conclusion: The suggested approach of discrete structure of dynamical system analysis allows to reduce the search areas of cyclic trajectories. We can apply properties of two described dynamical systems for all the considered n-dimensional models. Every other state transition diagram is isomorphic to one of diagrams of these systems.

References:

1. V.A. Likhoshvai et al. (2008) Theory of gene networks, In: Systems and computer biology, 395-480.
2. A.A. Akinshin, V.P. Golubyatnikov, and I.V. Golubyatnikov (2013) On some multidimensional models of gene network functioning. Journal of Applied and Industrial Mathematics, 7(3): 296-301.

CASE OF PROTIST RECENT DIVERGENCE RESOLVED USING PHYLOGENOMIC APPROACH

N.V. Annenkova
Limnological Institute SB RAS, Irkutsk, Russia
tasha.annenkova@gmail.com

Key words: phylogenomics, transcriptome, protists, dinoflagellates

Motivation and Aim: The taxonomy and evolution of protists, especially on the genus and species levels, is many times very difficult to unravel. One reason is that they often lack distinct morphological features making species delimitation and identification challenging (e.g. Glockner et al. 2013). and it is usually difficult to study their physiological features. Mechanisms of protist dispersal and speciation are also poorly understood compared to multicellular organisms. High throughput sequencing (HTS) should help a lot in understanding of speciation mechanisms within protists. HTS is widely used in gene expression analyzes and metagenomic, but its potential in evolutionary studies is also great. Phylogenomics can provide the drastic reduction in some negative factors, which influence on the tree topology.

The current study is focused on the dinoflagellates, which are a highly diverse group of protists. They inhabit both in freshwater and especially in marine ecosystems. Dinoflagellates have numerous unusual genetic features, in particular their genomes consist of 3–245 Gb of DNA held in from several to >100 chromosomes (Hou and Lin 2009). Recently, a clear example of recent radiation in free-living protists was observed among dinoflagellates (Logares et al. 2008, Annenkova et al. 2015). In particular, this group include marine species *Scrippsiella.* aff. *hangoei* and *Scrippsiella hangoei* and freshwater species *P. aciculiferum.* The first two have the same morphology, but differ in the rDNA and COB gene, and occur geographically very distant from each other (Antarctica and the Baltic Sea). The last one differs from *S. hangoei* in morphology, but is present in the same geographic area with S. hangoei. All three species are extremely similar in the neutral DNA markers. Phylogenetic analyses based on them tended to cluster *P. aciculiferum* with *S. hangoei* and left *S.* aff. *hangoei* out of this cluster but with low statistical support. We successfully used phylogenomic analyses based on transcriptomic data to resolve their relations.

Methods and Algorithms: Eight transcriptome data sets were obtained. Two originated from a strain of *Peridinium aciculiferum* and three each from *S. hangoei* and *S.* aff. h*angoei*. *De novo* assemblies were performed separately for each species using Trinity (Grabherr et al. 2011).

For the phylogenetic analysis an additional eight dinoflagellate species were chosen from publicly available databases based on the transcriptome availability and the relative closeness of the species to the studied dinoflagellates. Gene orthology searching was done using Proteinortho V. 5.06 (Lechner et al. 2011). Predicted peptide of all 11 transcriptomes grouped into a total of 48,218 orthologs. Groups, which contained genes from all species, were collected for further analysis. Each gene group was aligned and passed through a strict filtration of gaps and poorly aligned sites. Various numbers of the obtained alignments were concatenated and phylogenetic analyses were done using RaxML (Stamatakis 2014).

Results: The final data set contained 792 groups of ortholog genes (212,219 amino acid characters) from 11 species. Our result shown that *P. aciculiferum* diverged from the other two species (*S. hangoei* and *S.* aff. *hangoei*) before they split into two. The observation was congruent with the morphological identification of the species but not with data from neutral DNA markers. We suggested that the environmental features are more important factors of isolation for these dinoflagellates than geographic distance. It was also shown, that one should avoid using relatively small datasets: for us the tree topology constantly repeated when more than 200 genes were in alignment. The alternative topology within the hangoei/aciculiferum group occurred in several cases when the alignment contained 10 – 150 randomly picked genes. This can be associated with the recent origin of these species.

Conclusion: The work is one of the first phylogenomic studies of protists microevolution. Based on this study it is suggested the use of such complex data sets for phylogenetic analyses of protists, to resolve recently diverged species and avoid false positive signals. Transcriptomic data is a promising resource for phylogenomic analyses since sequencing of transcriptomes are cheaper and more easily obtained and assembled than sequencing whole genomes.

Acknowledgements: Author thanks Prof. Karin Rengefors and Dr. Dag Ahrén from Lund University (Sweden) for their help with study design and valuable advices.

References

1. Glockner, G., N. Hulsmann, M. Schleicher, A. A. Noegel, L. Eichinger, C. Gallinger, J. Pawlowski, R. Sierra, U. Euteneuer, L. Pillet, et al. (2013) The Genome of the Foraminiferan Reticulomyxa filose. Curr Biol. 24: 11-18.
2. Hou, Y., and S. Lin. (2009) Distinct gene number-genome size relationships for eukaryotes and non-eukaryotes: gene content estimation for dinoflagellate genomes. PLoS ONE 4: e6978.
3. Logares, R., N. Daugbjerg, A. Boltovskoy, A. Kremp, J. Laybourn-Parry J and K. Rengefors. (2008) Recent evolutionary diversification of a protist lineage. Environ. Microbiol. 10: 1231-1243.
4. Annenkova, N. V., G. Hansen, III Moestrup, and K. Rengefors. (2015) Recent adaptive radiation in a marine and freshwater dinoflagellate species flock. ISME J. doi:10.1038/ismej.2014.267
5. Grabherr, M. G., B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis , L. Fan, R. Raychowdhury, Q. Zeng, et al. (2011) Full-length transcriptome assembly from RNA-seq data without a reference genome. Nat Biotechnol. 29: 644-5.
7. Lechner, M., S. Sven FindeiЯ , L. Steiner, M. Marz, F. Stadler, and S. J. Prohaska. (2011) Proteinortho: Detection of (Co-)Orthologs in Large-Scale Analysis. BMC Bioinformatics 12: 124.
7. Stamatakis A. (2014) RAxML Version 8: A tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies". In Bioinformatics.

COMPARATIVE ANALYSIS OF SERUM PEPTIDES DETECTED IN SAMPLES FROM HEALTHY PERSONS AND COLORECTAL CANCER PATIENTS

I.V. Azarkin[1*], R.H. Ziganshin [1], S.I. Kovalchuk[1], G.P. Arapidi[1], O.M. Ivanova[1], V.O. Shender[1], N.A. Anikanov[1], V.M. Govorun [1,2], V.T. Ivanov[1]

[1]Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, RAS, Moscow, Russian Federation,
[2]Scientific Research Institute of Physical-Chemical Medicine, Moscow, Russian Federation
e-mail: garik.igor.azar@gmail.com
[*]Corresponding author

Key words: Colorectal cancer, Proteomics, Biomarkers, Metabolomics

Motivation and Aim: More than a million people a year worldwide develop colorectal cancer (CRC).Most of the CRC cases are sporadic, only 25% of the patients have a family history of the disease, and major genes causing syndromes predisposing to CRC only account for 5-6% of the total cases.

The aim of the present work was a search and identification of peptide markers of CRC in sera using modern mass spectrometry techniques.

Methods and Algorithms: Blood sera obtained from 50 patients with CRC and 50 healthy donors (control). Serum sample of each analyzed group were fractionated using magnetic beads with weak cation exchange surface, obtained eluates were analyzed by nanoLC-MS/MS using ABSciexTripleTOF 5600. All samples were analyzed by DDA (identification of serum peptides) and by SWATH (for label-free relative quantitative mass spectrometry analyses) approaches.

Results: As a result of LC-MS/MS analysis of sera more than 6000 unique peptides originated from the almost 1000 unique proteins were identified. Among identified peptides 786 were unique for CRC samples, and 125 of those were originated from the proteins unidentified in the control samples. For the control group there were 1075 unique peptides, 259 of which were originated from the proteins unidentified in CRC samples.

Conclusion: Our analysis allowed us to identify protein-protein interactions, responsible for various cellular processes, and to identify possible ways development of pathological states at molecular level.

Acknowledgements: This work was supported by the Russian Science Foundation (project No.14-50-00131).

ACTIVATION OF PROCASPASE-8 AT THE DISC

K. Bettermann, I.N. Lavrik*
Otto von Guericke University Magdeburg, Department of Translational Inflammation Research, Magdeburg, Germany
e-mail: inna.lavrik@med.ovgu.de
**Corresponding author*

Motivation and Aim: Systems biology is a rising field of science that connects mathematical modelling with powerful experimental methodology, maintaining a quantitative estimation of signalling pathways. Our group is working with systems biology to study death receptor networks. The death receptor family members are involved in controlling not only programmed cell death (PCD) but also proliferative pathways. PCD is indispensable for cellular homeostasis, such as elimination of unneeded, damaged, or infected cells in multicellular organisms. PCD imbalance leads to several severe diseases like cancer formation, neurodegenerative and autoimmune disorders.

In the last years the field of apoptosis were intensively studied and a huge amount of data were published but its quantitative regulation until recently has been poorly understood. By the help of mathematical models of death receptor signaling it was possible to achieve novel and crucial insights into the quantitative and mechanistic understanding of the death receptor network regulation.

We focus in our studies on one member of the death receptor family named CD95 (APO-1/Fas). The CD95 signalling pathway is characterized by the CD95 death-inducing signaling complex (DISC), composed of CD95, FADD, procaspase-8, procaspase-10, and c-FLIP, which plays a key role in apoptosis induction.

Recently, it was shown that procaspase-8 activation is mediated by death effector domain (DED) chains at the DISC. However, about their molecular architecture is little known.

Methods and Algorithms: We used different human cancer and T cell line models for quantitative mass spectrometry, Western blot and Immunoprecipitation assays and analysis and mathematical modeling to further unravel the molecular composition, dynamics and function of the DED chains.

Conclusion: By the combination of systems biology studies and biochemical methods it will be possible to receive novel data sets of the death receptor signalling, hereby focussing on the role of c-FLIP and caspase-8. Moreover, the studying of cancer and immune cells will provide not only an enhanced understanding but also offering a platform for drug development in the context of diseases associated to defects in death receptor signalling pathways.

DE NOVO TRANSCRIPTOME ASSEMBLY OF A SIBERIAN LARCH

V.V. Biryukov
Institute of Fundamental Biology and Biotechnology, Siberian Federal University, Krasnoyarsk, Russia
e-mail: vladbir2010@gmail.com

Keywords: Siberian larch, conifers, transcriptome assembly.

Motivation and Aim: Recent advances in next-generation sequencing (NGS) and the development of bioinformatic tools opened new opportunities in sequencing of large genomes. Besides, the sequencing and analysis of transcriptome of nonmodel organisms became available. On the contrary to the model organisms, conifers have some special features: slow growth rate, surprisingly high phenotypic plasticity, and large genome rich in repetitive sequences, mainly in the form of transposons and tandem repeats are among them. Due to these features, the process of the transcriptome assembly is quite complicated.

The study of the regulation of gene expression and cellular regulatory mechanisms is a sounding problem of modern biology. Here transcriptome assembly is an important part of that former. Therefore, the main aim of this study is transcriptome assembling of Siberian larch (*Larix sibirica* Ledeb.)

Methods and Algorithms: In the study, RNA-Seq data from buds of Siberian larch were assembled. Sequence reads were obtained on Illumina MiSeq platform in Laboratory of Forest Genomics at the Genome Research and Education Center in SibFU. Preprocessing of sequencing data was performed with Trimmomatic and FastQC tools. The *de novo* assembly was carried out using the Trinity transcriptome assembler, and the PRINSEQ and QUAST software was used to assess transcriptome assembly quality

Results: We obtained the primary transcriptome assembly of Siberian larch. The total length of all contigs is 26493048 bp. We mapped reads back to assembly contigs for assessing the quality of assembly (87.39% overall alignment rate, average and maximum coverage are 23 and 8014, respectively). Also, the basic metrics were evaluated: the total number of contigs is 43686, and 6919 of them are the contigs longer 1000 bp; the maximal contig length is 8512 bp; N50 is 878 bp.

Conclusion: We carried out the comparative analysis of the obtained transcriptome assembly to other transcriptomes [1]. Analysis shows a high similarity level between our contigs and the contigs studied previously. Thus, we have assembled about 70% of the total length of the transcriptome of Siberian larch, which can be used in further genome studies.

The study is a part of “Genomic studies major boreal coniferous forest tree species and their most dangerous pathogens in the Russian Federation” project funded by the Government of the Russian Federation (contract № 14.Y26.31.0004).

References:

1. J.L.Wegrzyn, J.M.Lee, B.R.Tearse, D.B.Neale. (2008) TreeGenes: A Forest Tree Genome Database, International Journal of Plant Genomics, 2008.

THE ROLE OF C-FLIP IN NF-KB ACTIVATION

M. Bollmann[1], J.H. Schmidt[1], N. Ivanisenko[2], V. Ivanisenko[2] and I.N. Lavrik[1*]
[1]Institute of Experimental Internal Medicine, Department Translational Inflammation Research, Otto von Guericke University Magdeburg, Germany
[2]Department of Systems Biology, Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
e-mail: inna.lavrik@med.ovgu.de
[*]Corresponding author

Key words: apoptosis, NF-κB pathway, computational modeling

Motivation and Aim: Apoptosis is the type of programed cell death which can be found in all multicellular organisms. In contrast to necrotic cell death, the apoptotic process is well controlled and essential to the development. The induction of apoptosis can be triggered by two different ways: the intrinsic pathway *via* mitochondria and the extrinsic pathway by a family of death receptors. One of these death receptors is represented by CD95 (APO-1/Fas), which induces after stimulation the formation of the death- inducing signaling complex (DISC). The DISC consists of CD95, FADD, procaspase-8 and -10 and c-FLIP. On the contrary the stimulation of the CD95 can also lead to 'pro-survival' responses. One of these responses is the activation of NF-κB, which is responsible for the transcription of genes for immune responses, apoptosis and cell growth. The mechanism of CD95-mediated NF-κB activation remains unknown and was in the focus of the current work.

Results and Conclusions: C-FLIP has been shown to play a role in the activation of NF-κB by binding to the regulatory subunit IKKγ (Nemo). Nemo is part of the IKK complex that phosphorylates IκBα, which is followed by ubiquitination and degradation of IκBα, and which; in turn, enables the NF-κB transcription factors to enter the nucleus and facilitate gene expression. In our studies we investigate the role of c-FLIP in CD95-induced NF-κB pathway with specific inhibitors designed using computational modeling. For the generation of the quantitative data of CD95-mediated NF-κB activation, we use western blotting, and/or imaging flow cytometry, which allows us to observe NF-κB activation on a single cell level. The implications of our findings will be discussed.

The work is supported by Russian Science Foundation 14-44-00011

References

1. Ozturk,S., Schleich,K., & Lavrik,I.N. Cellular FLICE-like inhibitory proteins (c-FLIPs): fine-tuners of life and death decisions. Exp. Cell Res. 318, 1324-1331 (2012).
2. Lavrik,I.N. & Krammer,P.H. Regulation of CD95/Fas signaling at the DISC. Cell Death. Differ. 19, 36-41 (2012)

ASSEMBLY AND ANNOTATION OF SIBERIAN LARCH CHLOROPLAST GENOME AND THE SEARCH FOR SINGLE NUCLEOTIDE POLYMORPHISMS

E.I. Bondar
Siberian Federal University, Krasnoyarsk, Russia
e-mail: bone-post@ya.ru

Key words: next-generation sequencing, de novo assembly, plastome, conifers.

Motivation: Cytoplasmic genomes of pine family species have a unique system of inheritance: mitochondrial DNA is inherited strictly through the maternal line and spreads with seeds while chloroplast DNA is inherited through the mail and extends through pollen. Currently, chloroplast DNA sequences of various species of coniferous (as well, as plants, in general) are the most important source of genetic markers to study the migration of paternal genes. Also it is one of the most popular objects for phylogenetic research. A number of sequences of chloroplast genomes of the pine family are published in NCBI Genbank. However, most of them belong to *Pinus* genus. There are only European and Western larch from the *Larix* genus with known chloroplast genomes.

Aim: The main aim of the study is assembling and annotation of chloroplast genome of Siberian larch (*Larix sibirica* Ledeb.), and identification of single nucleotide polymorphisms (SNPs) in these latter.

Methods and Algorithms: We used data of the whole genome sequencing of three Siberian larch trees from different regions from Urals, Krasnoyarsk, and Khakassia, respectively, for assembling. Sequence reads were obtained with Illumina HiSeq2000 in Laboratory of Forest Genomics at the Genome Research and Education Center in SFU.As a reference sequence were taken chloroplast genomes of *Larix decidua* Mill. [1] and *Larix occidentalis* Nutt. [2] from NCBI Genbank. The assembling has been done with the Bowtie2 mapping software and the SPAdes genomic assembler. The genome annotation was carried out using the Rapid Annotation using Subsystem Technology (RAST). For SNPs detection were used the Bowtie2 and UGENE software.

Results: The length of the final chloroplast genome was 122,561 bp, that is close to 122,474 bp in closely related *Larix decidua*. An annotation through the comparison with available data for *L. decidua* and *L. occidentalis* identified 121 coding regions, 34 among them represent RNA and 87 represent CDS genes. For three trees 13 SNPs were detected, 2 of them were found in the coding regions of the genome: *hypothetical protein* и *Cell division protein FtsH.*

Conclusion: The comparative analysis showed a strong similarity of the chloroplast genome of *L. Sibirica* and the chloroplast genome of *L. occidentalis*. The results of this research will be useful for further phylogenetic studies of conifers.

The study is a part of "Genomic studies major boreal coniferous forest tree species and their most dangerous pathogens in the Russian Federation" project funded by the Government of the Russian Federation (contract № 14.Y26.31.0004).

References:

1. C.S.Wu et al. (2011) Comparative chloroplast genomes of pinaceae: insights into the mechanism of diversified genomic organizations, *Genome Biol Evol,* **3**: 309-319.
2. M.Parks et al. (2009) Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes, *BMC Biol.,* **7:** 84.

APPLING BIOINFORMATICS METHODS FOR REVEALING MARKER PROTEINS SPECIFIC FOR ENDOTHELIAL DYSFUNCTION USING LC-MS/MS

A.G. Brzhozovsky*, L.Kh. Pastushkova, A.S. Kononichin, E.S. Tiise, I.M. Larina
Institute of Biomedical Problems RAS
**Corresponding author*

Key words: vascular endothelial dysfunction, marker proteins, liquid chromatography, mass spectrometry, associative network

Motivation and Aim: The main approach in a modern proteomic is using mass spectrometry methods. The most spread method is "shotgun". Due to this method one extracts proteins from the biological samples and then prepare via specific proteolysis enzymes (for example, tripsine). After that, we divide polypeptide chains using liquid chromatography (LC) and then we get mass spectrometry (MS) spectra. These spectra will be identified by us with special databases of amino sequences. After all this stages, we can use modern technical platforms and bioinformatics approaches to detect new marker proteins for drugs or for diagnostic of different diseases.

In our laboratory of Institute of Biomedical Problems we are trying to find biomarkers which are specific for endothelial dysfunction. The aim of our experiments is detection of marker proteins, which can characterize the early stage of vascular endothelial dysfunction that takes place in people with hypodynamia.

As we know, vascular endothelium plays a critical role in vascular tonus regulation via chemical mediators system. These mediators affect blood and vascular cells. Endothelial dysfunction is could be defined as one of the universal pathogenesis mechanisms of many diseases, including atherosclerosis, hypertensive disease, thrombosis etc. In this regard a hypertensive disease is aspect of modern lifestyle for peoples in developed countries. It is also negative factor for heart-vascular system, which includes lowering resistance of insulin and endothelial dysfunction.

Methods: Using proteomics methods based on mass- spectrometry we can reveal several proteins connected with endothelial function such as: endothelial receptor protein C, endothelin, thioredoxin (TRX), kallikrein-kinin system proteins (KNG1, KLK1). One of the projects aspects is using new development of "Waters" company – the Protein LynxGlobal Server v 2.3 (PLGS). It is a program pocket for LC-MS/MS processing. These programs can conduct quantitative and qualitative analysis.

Another important step of the analysis is reconstruction of associative nets between discovered proteins. On this stage we will used "STRing" and "ANDSyste" reconstructing systems. These systems are used to detect features of molecular-genes interaction. This interaction is realized in a certain protein groups or genes. We can use it to discover molecular mechanisms that underlie biological processes functioning. ANDSystem makes possible adding new proteins in reconstructing net system from global human proteins net. Thus, necessary proteins will be priorities.

A MODULE FOR INTEGRATING SBML-WRITTEN MATHEMATICAL MODELS INTO THE SOFTWARE PACKAGE «HAPLOID EVOLUTIONARY CONSTRUCTOR»

A.D. Chekantsev* , S.A. Lashin*
Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
Novosibirsk State University, Novosibirsk, Russia
e-mail: chekantsev@bionet.nsc.ru
**Corresponding author*

Motivation and Aim: Modeling the evolution of bacterial communities is an important task of modern computational biology. Haploid Evolutionary Constructor" (HEC) software package was developed in the ICG SB RAS to provide complex simulations of microbial (particularly bacterial) communities [1]. HEC allows users to simulate interactions between a population of cells and the environment, including a variety of evolutionary-genetic processes such as horizontal gene transfer between cells, loss of genes and mutations.

The interaction between populations realized by the processes of substrates production and consumption/utilization. In turn, the processes of synthesis and utilization of substrates are genetically determined. They are described by one or more gene networks. At present, many ready-made models focused on the processes of substrates metabolism are presented in databases and knowledge bases [2, 3, 4].

The aim of the current study is to integrate these well-known resources with the HEC via additional module supporting models import from existing databases (Biomodels, CellML) directly to the HEC.

Methods and Algorithms: In the HEC, we have implemented the multilayer modeling approach, where each "layer" represents a single computer model of biological objects and processes relating to a particular level of biological organization. In developing of the module used programming libraries libSBML, SOSlib and free cross-platform development environment Qt Creator.

Results: The implemented module allows us to extract the parameters and kinetic reactions formulas from the BioModels repository SBML models and to embed them into the HEC model (cellular metabolism layer).

Conclusion: Thus, this module allows you to use the "real world" models in HEC and to investigate the evolution and competition of gene networks in the context of relations between cellular metabolism and microbial community functioning

Acknowledgements: The study was supported by the RFBR 15-07-03879 grant.

References:

1. S.A. Lashin, Yu.G. Matushkin Haploid evolutionary constructor: new features and further challenges // In Silico. Biol. 2012. V.11. N. 3. P. 125-135.
2. N.Le Novиre, B. Bornstein, A. Broicher, M. Courtot, M. Donizelli, H. Dharuri, L. Li, H. Sauro, M. Schilstra, B. Shapiro, Jacky L. Snoep and M. Hucka BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems // Nucleic Acids Research. 2006. V. 36. P. 689-691.
3. I. Rojas**,** M. Golebiewski**,** R. Kania**,** O. Krebs**,** S. Mir**,** A. Weidemann and U. Wittig SABIO-RK: Integration and Curation of Reaction Kinetics Data. // Data integration in the life sciences. 2006. V. 4075. P. 94-103.
4. C. M. Lloyd, J. R. Lawson, P. J. Hunter and P. F. Nielsen The CellML Model Repository
5. // Bioinformatics. 2008. V. 24. I. 18. P. 2122-2123

COMPUTER ANALYSIS OF CLUSTERS OF TRANSCRIPTION FACTOR BINDING SITES BASED ON CHIP-SEQ DATA

A.I. Dergilev*, A.V. Svichkarev
Novosibirsk State University, Novosibirsk, Russia
Institute of Cytology and Genetics, Novosibirsk, Russia
e-mail: arturd1993@yandex.ru
** Corresponding author*

Keywords: ChIP-seq, binding sites, transcription factors, gene expression regulation

Motivation and Aim: The development of the methods for high-resolution mapping of transcription factor binding sites based on ChIP-seq (Chromatin ImmunoPrecipitation and sequencing) technique allows analyze the features of gene regulatory regions and chromatin structure at the genome scale [1]. ChIP-seq provides a qualitatively new data for the study of molecular mechanisms of coordinated regulation of the genes transcription. The objective is to study transcription factor binding sites (TFBS) in mouse genome using published coordinates from ChIP-seq experiments [2] and define clusters of sites of different protein transcription factors. Coordinates of transcription factor binding sites should be joined to clusters containing several different probably interacting sites in gene promoters and distal enhancers with following genomic annotation (location relative to target genes, promoters etc.). The task is to develop computer program for calculating clusters of different sites in the given genome by the coordinates of individual sites obtained experimentally using ChIP-seq.

Methods and algorithms: The sequencing data of 15 different transcription factors binding received in BED format (more than 100000 coordinates) were processed to define non-random clusters. A program written in C++ language calculates the distance between chromosomal coordinates in the mouse genome. The software produces a histogram of distances between binding sites of different factors (10, 20, 30 nt, etc., using user-defined scale) with visualization. Another program builds clusters of transcription factor binding sites with the output written to a text file in BED format.

Results: The tool developed has the following functions: Statistical analysis of ChIP-seq data; Processing gene BED files coordinates and gene lists; Counting of relative sites location in clusters; and Identification of gene ontology categories. Using this tool we confirmed cooperative localization Nanog, Oct4, Sox2 binding sites in clusters in mouse and human genomes based on latest ChIP-seq data.

Conclusion: Useful programs for operating with qualitatively new data were written. The output of clusters of different sites will be used for establishing complex signals and patterns of site location by the algorithm “Discovery” previously developed in the framework of the theory of data analysis (Data Mining and Knowledge Discovery).

Acknowledgments: This work was supported by RFBR (14-04-01906) and budget project ICG SB RAS. Computing was done at Shared Facility Center “Bioinformatics”.

References:

1. E.V. Ignatieva et al. (2015) Regulatory genomics: combined experimental and computational approaches, *Russian Journal of Genetics*, V. 51(4), P. 334–352.
2. X. Chen et al. (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells, *Cell*, 133(6), P. 1106-17.

THE DEVELOPMENT OF NEW PHOTOACTIVATABLE MOLECULAR SWITCHES FOR THE TREATMENT OF EYE DISEASES CAUSED BY DEGENERATION OF THE RETINA

L.N. Dvoretckaia
St Petersburg Academic University, Saint-Petersburg, Russia
e-mail: Liy28@rambler.ru

Key words: retina, rhodopsin, quantum chemistry, molecular modeling

Motivation and Aim: Diseases associated with degeneration of the retina, such as age maculodestrophy and retinitis pigmentosa, widely distributed diseases for which has still not found effective drugs. This project aimed at solving the fundamental problem of creating substances with predetermined biological and photochemical properties, which allow to restore visual function in retinal lesions. The mechanism of action of these drugs is based on selective compounds blocking voltage-dependent potassium channels.

Methods and Algorithms:

1. Generation of libraries of functional derivatives on the based on the known type (azobenzene, spiropyran, diariletan, fulgides, indanilidenpirrol) molecular switches
2. The selection of the most promising candidates through rational design by computer modeling methods, and quantum chemistry (molecular docking on the basis of the crystal structure of voltage-dependent potassium channel, molecular dynamics, DFT, TD-DFT , CASPT2, SORCI)

Results: There are currently we did molecular design, computer modeling, the study of the photochemical cycle and been developed a series of strategies for the synthesis of molecular switches based on indanilidenpirrola, simulating photochemical properties of rhodopsin.

Conclusion and Availability: The compounds of based indanilidenpirrola are more promising than with the proposed based on azobenzene candidates for use as molecular switches for restoring visual function of the affected retina.

References:

1. Sinicropi A, Martin E, Ryazantsev (2010) MN et.al. PNAS.
2. Ryazantsev M. N, Altun A., Morokuma K. (2012) Color Tuning in Rhodopsins: The Origin of the Spectral Shift between the Chloride-bound and Anion-free Forms of Halorhodopsin // J. Am. Chem. Soc, Vol. 134, Issue 12. P. 5520-5523.

ABOUT POPULATION GENOMICS OF SOME ANIMAL SPECIES

V.S. Fedotova*, M.G. Sadovsky, Y.A. Putintseva
Siberian Federal University, Krasnoyarsk, Russia
e-mail: vasilinyushechka@gmail.com
**Corresponding author*

Key words: mitochondrial genomes, clustering, evolution

In this work population diversity has been researched by genomic methods, specifically mitochondrial genome clustering of different organism groups. All mitochondrial genomes deposited in EMBL-bank database have been studied (http://www.ebi.ac.uk/ena), statistical stability of data was defined by linear method (k-means clustering) also data was clustered by nonlinear method (Elastic map).

Source database of mitochondrial genomes contains variety of species in different genuses: for example, the genus *Equus* contains 257 genomes of species, but the genus *Ablennes* contains only 1 genome. The amount of genomes in source database is 3726. Stability of classification has studied by k-means clustering method depending on the completeness of the different genera. For this work used two databases: source database which received from the bank and edited database that contained no more than 50 species in the same genus. This limitation was determined by the structure of the source database of genomes. The amount of genomes in edited database is 2989.

Dividing plurality of genomes has studied using k-means clustering into two classes. Each genome is represented by a point in the 63-dimensional space frequencies of triplets. Stable groups of genomes were sorted out, always applied to the same class in 500 implementations out of 500 while being divided.

By classes distribution the symmetrical difference was discerned for every completely stable group. Moreover, there are 71% absolute group genomes in source database and 74% in edited database out of overall quantity of genomes. The powerfulness of genome multitude was formed symmetrical difference of edited database has decreased comparing with the source database. Static stability while classification modeling was defined by k-means clustering method.

Elastic maps provide a tool for nonlinear dimensionality reduction. By their construction, they are a system of elastic springs embedded in the data space. This system approximates a low-dimensional manifold. [1]

For clustering by elastic map method was taken only one type of animal, it is *Chordata*. This type contains five classes in the database. They are *Actinopterygii*, *Amphibia*, *Aves*, *Mammalia*, *Reptilia*. Detailed map was constructed.

The class *Actinopterygii* formed two clusters, *Amphibia* divided into two clusters, *Aves* formed one big cluster. The class *Mammalia* divided into four clusters, also as the class Reptilia.

The division in classes has been uneven. Now let's review the most interesting distributions of the class *Mammalia*. Mostly all representatives of the *Bears* were divided in a single cluster, except *Short-faced bear*, it got in the biggest cluster. All representatives of families *Hominidae* and *Gibbons* got in a separate cluster. All representatives of *Deers* и *Oceanic dolphins* got in the biggest cluster. Representatives of families *Old World monkeys*, *Muridae*, *Shrews*, *Talpidae, Rorquals, Bovidae, Leaf-nosed bats*, *Cricetidae, Elephantidae, Lemuridae* were divided between two clusters. All *Rhinoceros*, *Eared seals,* and *Earless seals* were divided in the last cluster.

Family *Crocodylidae* from the class *Reptilia* was divided, two genera form separate cluster. *Agamidae*, *Gekkonidae*, *Chamaeleonidae* and different genera *Turtles* and *Tortoises* was divided between two clusters. Cluster distribution of the *Amphibia* class: there are *Bombinatoridae*, *Bufonidae*, *Microhylidae* and another close families in the first cluster. There is only one family similar to the first cluster by composition, it is *Pipidae*. Another representatives of the second cluster are *Salamanders*, *Caecilians*, *Ambystomatidae*, etc. A large number of species that are not related to any clusters mainly they are representatives of the family *Salamandridae*.

Database reduction result in increase of clustering stability. Clustering of genomes based on the triplet structure by k-means clustering gives results generally appropriate taxonomic idea of the type *Chordata*, but also reveals an interesting deviation from the standard taxonomy, which is expected to research in further work.

References:

1. A. N. Gorban, A. Y. Zinovyev, Principal Graphs and Manifolds, In: Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods and Techniques, Olivas E.S. et al Eds. Information Science Reference, IGI Global: Hershey, PA, USA, 2009. 28–59.

THE ABSCISIC ACID EFFECT ON THE PECTOBACTERIUM ATROSEPTICUM TRANSCRIPTOME PROFILE

R.F. Gubaev[1,2,*], V.Y. Gorshkov[1,2], O.E. Petrova[1], A.G. Daminova[1] Y.V. Gogolev[1]
[1]*Kazan Institute of Biochemistry and Biophysics RAS, Kazan, Russia*
[2]*Kazan Federal University, Kazan, Russia*
e-mail: rimhuman@gmail.com
[*]*Corresponding author*

Key words: plant-pathogen interactions, abscisic acid, next generation sequencing, Pectobacterium, virulence

Motivation and Aim. The mutual "eavesdropping" of partner's endogenic signals is a common phenomenon in plant-pathogen interactions. The host plant-derived regulatory molecules could serve as signals or metabolism affecting factors for the pathogen. Abscisic acid (ABA) is a plant hormone that regulates responses to the abiotic stressors, water status and ontogenetic processes in plants. The host-plant's water status controlled by ABA is an important factor affecting virulence of the phytopathogen *Pectobacterium atrosepicum* (*Pba*). It is well-known fact that *Pba*-induced disease symptoms are not expressed under low humidity when the concentration of ABA is elevated in plants. However, there are still no data about ABA impact on the pathogen's physiology. Previously obtained data has shown that ABA could be a negative regulator of Pba virulence because this phytohormone reduces activity of pectatelyases -the powerful ammunition of Pba. The aim of the present study was to explore the ABA effects on the pathogen's transcriptomic profile for more precise analysis of the physiological changes.

Methods and Algorithms. Pba cell cultures were incubated during 24 hours in D7 media containing pectin as a carbon source in the presence or absence of 0.1 mM ABA. Total RNA was extracted using the Extract RNA reagent kit (Evrogen). DNA was depleted using TURBO DNA-free kit (Ambion). The cDNA library was prepared using the NEB Next Ultra RNA Library Prep Kit (NEW ENGLAND BioLabs). The reads were obtained using the MiSeq next generation sequencing platform. The obtained reads were trimmed and quality sorted using Trimmomatic. Processed reads were aligned along the P*ba* genome using Bowtie2 aligner. Read counts were estimated with Cufflinks using RPKM metrics. The differential gene expression was analyzed using the Cuffdiff package.

Results. The addition of abscisic acid to *Pba* cell cultures resulted in changes in the transcriptomic profile of the pathogen. The expression of many *Pba* virulence-related genes, including those that encode effector proteins of type III secretion system, pectate lyases, regulatory and transport proteins, was shown to be ABA-repressed.

Conclusion. The negative impact of ABA on virulence-related gene expression profile was demonstrated. Thus ABA impact on the plant pathogen physiology is complexed: it could influence at the virulence exhibition (i.e. pectolytic enzymes activity) and the virulence related gene expression (i.e. type III secretion effector proteins). These facts indicate that ABA could play a multiple role in plant physiological state related with both biotic and abiotic stress factor responses.

Acknowledgements. This study was supported by the RFBR, research project No. 14-04-01750_A. and RUSSIAN SCIENCE FOUNDATIONRESEARCH PROJECT #15-14-10022

THE ABSCISIC ACID EFFECT ON THE PECTOBACTERIUM ATROSEPTICUM TRANSCRIPTOME PROFILE

R.F. Gubaev[1,2,*], V.Y. Gorshkov[1,2], O.E. Petrova[1], A.G. Daminova[1] Y.V. Gogolev[1]
[1]*Kazan Institute of Biochemistry and Biophysics RAS, Kazan, Russia*
[2]*Kazan Federal University, Kazan, Russia*
e-mail: rimhuman@gmail.com
[*]*Corresponding author*

Key words: plant-pathogen interactions, abscisic acid, next generation sequencing, Pectobacterium, virulence

Motivation and Aim. The mutual "eavesdropping" of partner's endogenic signals is a common phenomenon in plant-pathogen interactions. The host plant-derived regulatory molecules could serve as signals or metabolism affecting factors for the pathogen. Abscisic acid (ABA) is a plant hormone that regulates responses to the abiotic stressors, water status and ontogenetic processes in plants. The host-plant's water status controlled by ABA is an important factor affecting virulence of the phytopathogen *Pectobacterium atrosepicum* (*Pba*). It is well-known fact that *Pba*-induced disease symptoms are not expressed under low humidity when the concentration of ABA is elevated in plants. However, there are still no data about ABA impact on the pathogen's physiology. Previously obtained data has shown that ABA could be a negative regulator of Pba virulence because this phytohormone reduces activity of pectatelyases -the powerful ammunition of Pba. The aim of the present study was to explore the ABA effects on the pathogen's transcriptomic profile for more precise analysis of the physiological changes.

Methods and Algorithms. Pba cell cultures were incubated during 24 hours in D7 media containing pectin as a carbon source in the presence or absence of 0.1 mM ABA. Total RNA was extracted using the Extract RNA reagent kit (Evrogen). DNA was depleted using TURBO DNA-free kit (Ambion). The cDNA library was prepared using the NEB Next Ultra RNA Library Prep Kit (NEW ENGLAND BioLabs). The reads were obtained using the MiSeq next generation sequencing platform. The obtained reads were trimmed and quality sorted using Trimmomatic. Processed reads were aligned along the P*ba* genome using Bowtie2 aligner. Read counts were estimated with Cufflinks using RPKM metrics. The differential gene expression was analyzed using the Cuffdiff package.

Results. The addition of abscisic acid to *Pba* cell cultures resulted in changes in the transcriptomic profile of the pathogen. The expression of many *Pba* virulence-related genes, including those that encode effector proteins of type III secretion system, pectate lyases, regulatory and transport proteins, was shown to be ABA-repressed.

Conclusion. The negative impact of ABA on virulence-related gene expression profile was demonstrated. Thus ABA impact on the plant pathogen physiology is complexed: it could influence at the virulence exhibition (i.e. pectolytic enzymes activity) and the virulence related gene expression (i.e. type III secretion effector proteins). These facts indicate that ABA could play a multiple role in plant physiological state related with both biotic and abiotic stress factor responses.

Acknowledgements. This study was supported by the RFBR, research project No. 14-04-01750_A. and RUSSIAN SCIENCE FOUNDATIONRESEARCH PROJECT #15-14-10022

SMALL MOLECULES AS MODULATORS OF LIFE/DEATH DECISIONS AT TARGET PROTEINS OF THE DISC

L. Hillert[1], C. Arndt[1], S. Pietkiewicz[1], N. Ivanisenko[2], V. Ivanisenko[2] and I.N. Lavrik[1*]
[1]*Institute of Experimental Internal Medicine, Department Translational Inflammation Research, Otto von Guericke University Magdeburg, Germany*
[2]*Department of Systems Biology, Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia*
e-mail: inna.lavrik@med.ovgu.de
[*]*Corresponding author*

Motivation and aim

The programmed cell death, also called as apoptosis, is an essential cellular process in all multicellular organisms and important for cell homoeostasis. A dysregulation of apoptosis is linked to cancer, autoimmune diseases and neurodegeneration. Apoptosis can be induced by an intrinsic or an extrinsic pathway, the latter is induced by stimulation of the death receptor (DR) at the cell surface. One of the best-characterized DRs is CD95, which is also known as APO-1 or Fas. Stimulation of the CD95 by its ligand leads to the formation of the death-inducing signaling complex (DISC), comprising CD95, FADD, c-FLIP, procaspase-8, and procaspase-10. After formation of the DISC procaspase-8 is activated by autocatalytic processing, leading to the initiation of the caspase cascade and, finally, cell death. The activity of caspase-8 at the DISC is regulated by c-FLIP, thereby influencing pro- and anti-apoptotic signaling pathways positively or negatively. In recent years small molecules have been discovered and developed as molecularly targeted drugs to personalize treatments of patients. The development of targets for the CD95 DISC and thereby of DR-based new therapies was the aim of this work.

Results and Conclusions

The inhibition of the target proteins is triggered by interaction of the small molecule to specific functional domains of the protein. In cooperation with Novosibirsk Institute of Cytology and Genetics different potential small molecules based on the crystal structure of FADD and c-FLIP were found by bioinformatical screening and modeling. Binding of inhibitors of FADD (FADDins) and inhibitors of FLIP (FLIPins) to hydrophobic pockets shall allow to change protein complex formation and protein activity and thereby to modulate pro- and anti-apoptotic signaling pathways.

In our studies the effect of the small molecules targeting the proteins FADD and FLIP are investigated. The effects of these small molecules on CD95-mediated signaling pathways and protein complex formation are analyzed in detail by biophysical, cellular and cellular biology methods. In conclusion these studies should provide new insights into molecular mechanisms of DR signaling and paving the road towards development of new treatments in personalized medicine.

The work is supported by Russian Science Foundation 14-44-00011

MICROBIAL DIVERSITY IN ZHOIGAN HOT SPRING (EASTERN SAYAN, RUSSIA)

E.S. Kashkak[1*], N.L. Belkova[2]
[1]*Baikal Institute of rational nature management SB RAS, Ulan-Ude, Russia*
[2]*Limnological Institute SB RAS, Irkutsk, Russia*
e-mail: klslena@yandex.ru
**Corresponding author*

Key words: hot spring, metagenomic analysis, diversity, Eastern Sayan, Proteobacteria, Firmicutes, Bacteroidetes

Motivation and Aim: The first studies on thermophilic microorganisms from hot springs were focused on isolation and characterization of thermophilic strains using culture-dependent approaches. However, ~99% of the microorganisms within a particular environment proved uncultivable [1]. Molecular phylogenetic methods, based on the comparison of the sequences of the 16S rDNA, have revealed the hidden diversity of microorganisms from hot springs water and microbial mats. Metagenomic approaches have been extensively used to describe microbial communities and identify novel thermophilic microorganisms. The aim of this study was structure and composition analysis of microbial communities of water and microbial mat of Zhoigan mineral springs by high-throughput sequencing of 16S rDNA amplicons.

Materials and methods: The territory of Zhoigan mineral springs is located in the Eastern Sayan Mountain, at an altitude of 1550 meters above sea level (N52 36.99 E099 00.467) and contained more than 20 outlets of hot (up to 40°C) and cold (below 20°C) water. The carbonated springs are located on the travertine field where mineral water is poured and microbial mats are formed, which have been the objects of study.

The physical and chemical characteristics of the water at the sampling sites were carried out by conventional methods [2]. DNA was extracted with commercial kits a DNA-sorb (AmpliSens, Moscow) and Bacterial Genomic DNA isolation (Axygen, USA), according to the modified methods [3]. The V3-V4 region of the 16S rRNA genes was amplified with the primer pair 343F (5'-CTCCTACGGRRSGCAGCAG-3') and 806R (5'-GGACTACNVGGGTWTCTAAT-3') combined with Illumina adapter sequences. Metagenomic analysis of amplicons was conducted on MiSeq Illumina sequencer at the SB RAS Genomics Core Facility (ICBFM SB RAS, Novosibirsk, Russia).The following resources was used for bioinformatics of obtained libraries: RDP Pipeline (http://pyro.cme.msu.edu/), EzTaxon [4] and Mothur v.1.22.0 (http://www.mothur.org).

Results and discussion: Water and microbial mats were taken on the travertine field of 30 m from the thermal output spout, where the maximum accumulation of biomass of microbial mats was observed. The water temperature and pH value were 38°C and 7.7, respectively. Total mineralization amounted to 900 mg/L. The content of dissolved CO_2 was 352.0 mg/L, whereas total iron concentration – 4.2 mg/L.

In total, 86113 and 33060 valid nucleotide sequences were retrieved from the two libraries through metagenomic sequencing of water and microbial mats of Zhoigan mineral spring, respectively.

In the water 86111 sequences were attributed to the domain Bacteria and 2 sequences – to Archaea. Thus, Archaea are less than 0.1% of the water spring microbial community. In microbial mats Achaea were not found. In microbial mats, all identified

sequences were attributed to the domain Bacteria. The dominant bacterial groups in these springs are characteristic for neutral or slightly alkaline springs with moderately high temperature. The presence of the minor groups of Archaea (less than 0.1%) is because they are mainly dominant component in the extreme habitats with high temperature, high acidity or alkalinity [5].

Bacterial community of water was characterized by high species diversity where 21 phyla were presented. The representatives of phyla Proteobacteria (73.4%) and Bacteroidetes (16.4%) were dominated in the microbial community. At a lower taxonomic level the representatives of *Hydrogenophaga* (11%, Betaproteobacteria), *Acidovorax* (3%, Betaproteobacteria), *Serratia* (2.8%, Gammaproteobacteria), *Sediminibacterium* (2.1%, Sphingobacteria), and *Perlucidibaca* (2%, Gammaproteobacteria) were determined as dominant genera. Less than 0.2% of sequences were assigned to phyla with uncultivated genera. Archaea were attributed to methane-producing bacteria of the genus *Methanospirillum* (class Methanomicrobia, phylum Euryarchaeota).

In microbial mat three phyla: Proteobacteria (36.5%), Firmicutes (20.8%) and Bacteroidetes (16.4%) were dominated. Proteobacteria includes taxa of five classes from which alpha, beta and delta-proteobacteria are prevailed. In microbial mats among the Proteobacteria anaerobic bacteria of the genus *Geobacter* (1.9%, Deltaproteobacteria) with capacity to oxidize the organic substances as well as metals, including iron, and hydrogen-oxidizing bacteria of the genus *Hydrogenophaga* (2%, Betaproteobacteria) were predominated. Among the other phyla anaerobic acetate- producing bacteria of the genus *Acetoanaerobium* (1.95%, Clostridia), filamentous anaerobic thermophilic bacteria of the genus *Caldilinea* (2%, Caldilineae), facultative anaerobic and obligate organotrophic bacteria of the genus *Ignavibacterium* (2.2%, Ignavibacteria) were dominated. About 0.1% of the microorganisms referred to phylum with non-cultivated representatives.

Conclusion: Firstly, the results of metagenomic analysis of microbial communities of water and microbial mats of Eastern Sayan mineral springs show considerable diversity of prokaryotes. Proteobacteria and Bacteroidetes were dominated in the community of water sample whereas Proteobacteria, Firmicutes, and Bacteroidetes are prevailed in microbial mats of the springs. Methanogens of genus *Metanospirillum* (less than 0.1%, Euryarchaeota) relating to Archaea were found only in the water of mineral spring.

Acknowledgements: This work was supported by the Ministry of Education and Science of the Russian Federation (Project No. 1990) and Russian Foundation for Basic Research (RFBR) grant №15-04-01275a, №15-44-04335-r_sibir_a.

References:

1. O. Lypez-Lypez et al. (2013) Hot Spring Metagenomics. Life, 2: 308-320; doi:10.3390/life3020308.
2. B.B. Namsaraev et al. (2005) Field workshop on water microbiology and hydrochemistry: 68.
3. N.L. Belkova (2009) Molecular genetic methods of analysis of microbial communities // In: The diversity of microbial communities in inland waters Russia, A.M. Andreeva (Eds.), 53–63 (Publisher).
4. J. Chun et al. (2007) EzTaxon: a web-based tool for the identification of prokaryotes based on 16S ribosomal RNA gene sequences, Int. J. Syst. Evol. Microbiol., V. 57. Pt.10: 2259–2261.
5. V.M. Gumerov et al. (2011) Molecular analysis of microbial diversity in the Zavarzin Spring, Uzon Caldera, Kamchatka, Microbiology, 80: 258-265.

COMPARATIVE TRANSCRIPTOME ANALYSIS OF LIFE FORMS OF THE MOSS *PHYSCOMITRELLA PATENS*

R.A. Khazigaleeva*, I.A. Fesenko, E.S. Kostrukova, T.A. Semashko, G.P. Arapidi, K.A. Babalyan, V.M. Govorun
Shemyakin&Ovchinnikov Institute of bioorganic chemistry RAS, Moscow
e-mail: khazigaleeva.regina@gmail.com
**Corresponding author*

Key words: comparative analysis, protoplasts transcriptome, gene expression, gene ontology

Motivation and Aim: The comparative analysis of transcription can be used for studies of gene expression and the function, comparing the biochemical pathways, identification of functional homologues. In our work, a comparative analysis of two transcription of life forms, as well as the moss *Physcomitrella patens* protoplasts to identify genes differentially expressed (DE) in different cell types.

Methods and Algorithms: The transcriptome were analyzed in several biological and technical replicates using genomic SOLID analyzer 4. We obtained 173, 197 and 204 million reads for gametophore, protonema and protoplast samples, respectively. The number of uniquely mapped filtered reads was 31, 36 and 38 million for gametophore, protonema and protoplast samples, respectively. According to the summary, the moss *P. patens* contains 33096 coding region (CDS). We found the expression of 16201 CDS (more than one read per million).To validate the accuracy and to evaluate the distortion that occurred during library preparation, the transcriptional levels of seventeen genes were analyzed by quantitative real-time PCR (qRT-PCR).The Spearman correlation values of gene expression obtained by qRT-PCR and RNA-seq methods was 0.7, 0.7 and 0.8 for gametophore, protonema and protoplast samples, respectively. We used the median value of different repeats of RNA-seq for each sample.

Results and Conclusions: In a comparative analysis of two life forms moss - and protonemal gametophores was discovered 3170 DE genes of which the expression was increased in 1551 and decreased in 1619. When allocating protoplast cells moss *P. patens* experience severe stress associated with exposure to the enzyme preparation drayzelaza and loss of the cell wall. The comparative analysis of protonema and protoplasts, we have identified 1788 DE genes, while for 1376 increased expression in protoplasts. Analysis of Gene Ontology (GO) terms showed that in protoplasts the transcriptional level of genes involved in responses to different stress factors increased, e.g. the response to abiotic stimulus (GO:0009628), response to cold (GO:0009409), response to temperature stimulus (GO:0009266), response to oxidative stress (GO:0006979). Severe stress in protoplasts could also be detected by the increased expression of genes participating in protection from reactive oxygen species (ROS), such as Pp1s152_86V6, putative ascorbate peroxidase 1.

Notably, it also increased the level of transcription of genes encoding enzymes involved in the synthesis of jasmonic acid, one of the stress hormones synthesized by biotic stress.

Acknowledgements:
This work was supported by the Russian Science Foundation (project No. 14-50-00131)

THE DNA NITROUS BASES OSCILLATIONS UNDER THE ACTION OF THE EXTERNAL PERIODIC FORSE

L.A. Krasnobaeva*[1,2], L.V. Yakushevich[3]
[1]*Siberian State Medical University, Tomsk, Russia*
[2]*Tomsk State University, Tomsk, Russia*
[3]*Institute of Cell Biophysics RAS, Pushchino, Moscow region, Russia*
e-mail: kla1983@mail.ru
[*]*Corresponding author*

Key words: DNA dynamics, forced angular oscillations of nitrous bases , modeling, interferon alpha 17 (IFNA17)

Motivation and Aim: The dynamical properties of different DNA sequences as well as the connection between their dynamics and function are still poorly studied. In this paper we investigate the angular oscillations of nitrogenous bases under the action of external periodic force.

Methods and Algorithms: As an object of the research we selected the sequence of the gene coding interferon alpha 17 (IFNA17), that is known as antiviral and antitumor agents. As a mathematical model, we used a system of two coupled nonlinear differential equations that imitated free angular oscillations of nitrous bases in both polynucleotide chains and took into account the effects of dissipation [1]. The equations were complemented by the terms imitating the action of external periodic force. To find the solutions of the model equations, a variety of analytical methods, including the average field approximation, the energetic method [2] and the method of concentrations [3] were used.

Results: In the linear approximation, we obtained general solution of the problem, constructed the dispersion curves and found the conditions of resonance. In the nonlinear case, we found analytical formulas which describe the dependence of the coordinate and the velocity of the nonlinear conformational waves – kinks, on time for different values of the amplitude (F_0) and the frequency (Ω) of the external generalized force .

References:

1. L.V. Krasnobaeva, L.V. Yakushevich (2015) Rotational dynamics of bases in the gene coding interferon alpha 17 (IFNA17), *Journal of Bioinformatics and Computational Biology*, **13: 1,** 1540002 .
2. D.W. McLaughlin, A.C. Scott (1978) Perturbation analysis of fluxon dynamics, *Phys. Rev. A*, **18:** 1652-1660.
3. L.V. Krasnobaeva, L.V. Yakushevich (2008) A new approach to studies of non-linear dynamics of kinks activated in inhomogeneous polynucleotide chains, *International Journal of Nonlinear Mechanics*, **43:** №10, 1074-1081.

HOW MANY SPECIES OF CELANDINE (*CHELIDONIUM*, PAPAVERACEAE)?

D.A. Krivenko[1]*, O.S. Dymshakova[2]
[1]*Siberian Institute of Plant Physiology and Biochemistry SB RAS, Irkutsk, Russia*
National Research Irkutsk State Technical University, Irkutsk, Russia
[2]*Institute of Plant and Animal Ecology UB RAS, Ekaterinburg, Russia*
e-mail: krivenko.irk@gmail.com
**Corresponding author*

Key words: celandine, cryptic species, chloroplast fragment, nucleotide alignment, phylogenetic tree

Motivation and Aim: Chelidonium majus L. sensu lato (Papaveraceae Juss.) – celandine – morphological and karyological polymorphic species. Aim of the study is to identify cryptic species through detailed taxonomic boundaries of the *Ch. majus* and differentiate the populations based on genetic and caryological data.

Methods and Algorithms: We analyzed the genomic and chloroplast fragments of 36 specimens of celandine: internal transcribed spacers, ITS1 – 5.8S – ITS2 and spacers within the chloroplast region trnL – trnF. All genome fragments were amplified in PCR and sequenced. Obtained sequences were searched for close homologues with help of Basic Local Alignment Search Tool (NCBI/BLAST) and analyzed in BioEdit 7.0.9.0 program. The genetic similarity of ITS was 98.2 – 100 % while chloroplast fragment was more conservative – 99.6 – 100%. Only 10 of 36 sequences had single substitutions in chloroplast fragment. Both fragments except genes for 5.8S rRNA and trn were combined to construct alignment for phylogenetic analysis. The Bayesian phylogenetic tree was reconstructed on the basis 1088 bp nucleotide alignment. Were also counted of the chromosome number ($2n$) in samples of celandine included in genetic analyze. The chromosome counts were carried out using pressed temporary preparations of dividing cells (at metaphase) from the apical portion of roots by standard method.

Results and Conclusions: There were three well separated clades (*A*, *B*, *C*) on the phylogenetic tree. Clades A and B had a common ancestor and separate celandine from two areas West Siberia – Central Asia (clade *A*) and Europe – Caucasus (clade *B*) with chromosome number $2n = 12$. The clade *C* which was most distinct from others and includes specimens from Far East with $2n = 10$ and south part of East Siberia with $2n = 11$ and rarely detected samples with $2n = 12$. We assume that the clade *A* and *B* are independent species in which $2n = 12$. Also the clade *C* combines 2 species: one species with $2n = 10$, and a hybrid with $2n = 11$ or 12.

Cryptic species represent a significant gap in the study of biodiversity and taxonomy questions. Identification and study of such species is very important to disclose these issues to the knowledge of evolution and speciation, for the protection of rare and economically valuable species.

Acknowledgements: We are grateful to N.V. Kulakova for help in the treatment of molecular genetic analysis. And also S.G. Kazavonsky and A.V. Verkhozina for providing samples of celandine. The reported study was partially supported by RFBR, research projects No. 14-44-04105-a_sibir, 14-47-04125-a_sibir and 15-04-05372-a.

COMPUTER ANALYSIS OF GENE LOCATION RELATIVE TO CHROMOSOME CONTACTS REVEALED BY CHIA-PET AND HI-C

E.V. Kulakova[1,2*], N.R. Battulin[1,2]
[1]*Nosobirsk State University, Novosibirsk, Russia*
[2]*Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia*
e-mail: kylakovaekaterina@gmail.com
[*]*Corresponding author*

Key words: sequencing, ChIA-PET, Hi-C, chromosome contacts, genome, CTCF sites

Motivation and Aim: Several technologies based on chromatin immunoprecipitation (ChIP) have been developed to study the binding of transcription factors (TF) to genomic DNA including microarray (ChIP-chip), ChIP-PET and ChIP-Seq [1]. Such methods are unable to determine the target genes of the distal TF binding sites that present a challenge for bioinformatics research. Another challenge is to define whether such distal binding sites are functional, i.e. physically proximal to target gene promoters via chromosome loops attracting RNA polymerase II complex for gene transcription [2,3]. Therefore, identification of genome-wide distal chromatin interactions that lead the regulatory elements to their target genes may provide novel insights into the study of transcription regulation. Chromatin Interaction Analysis with Paired-End-Tag sequencing (ChIA-PET) method fits these demands still requiring development of specialized high-throughput software for data integration and statistical estimation of the data obtained [2,3]. The aim of the work was to develop a computer program for statistical data analysis and test it on CTCF (CCCTC-binding factor) binding sites, genes and spatial topological domains.

For Hi-C data [4-6], at the first stage, as in the ChIP-seq method, treatment with formaldehyde is conducted, resulting in the formation of protein-nucleic crosslinking and the extraction of complexes with antibodies against a distinct nuclear protein. Fragmentation is performed with restriction endonucleases, resulting in the formation of significantly long fragments, and followed by their self-ligation under strong dilution conditions until complex cleavage. These conditions provide ligation of DNA molecule ends, which appear to be in close spatial proximity. Distinct DNA molecules are subsequently extracted (usually via biotin labeling) and processed for sequencing. It is followed by a computer alignment of sequenced DNA fragments formed by a pair of genomically distant but spatially proximal regions, which appeared to comprise the same sequenced DNA fragment as a result of a procedure described above. These regions are identified as the chromosome contact regions.

The study of chromosomal contacts raises the problem of statistical analysis, which requires the processing of a substantial amount of experimental data that exceeds the standard volume of ChIP-seq data. The presence of chromosomal territories has been confirmed via the Hi-C method. Interchromosomal contacts were shown to be highly dynamic [6].

Methods and Algorithms: We used data on the spatial domains in the genome of the mouse embryonic stem cells and in the human genome, data on the location of CTCF binding sites clusters obtained by ChIA-PET as well as obtained experimentally by methods ChIP-seq, Hi-C, ChIA-PET [2]. Gene annotation was obtained from UCSC Genome Browser (http://genome.ucsc.edu). The processing of Hi-C data includes the isolation of the spatial chromosomal domains; such information is stored in databases available via Internet (for instance, 3DGD and Mouse Encode Project at B.Ren Lab -

http://chromosome.sdsc.edu/mouse/). We developed computer program for processing of these data.

Results: The result of the calculation is a distribution of CTCF transcription factor binding sites on domains on the human chromosomes. The distributions of human genes relative CTCF binding sites and a randomly generated list of such sites as the program output were used to estimate statistical significance of the associations found.

For Hi-C data analysis in mouse [7] we used gene ontology (GO) analysis of genes located on the borders of topological genome domains. We show enrichment of GO categories related to cell membrane and extracellular function that indicate to location of chromosome domains in interphase nuclei close to cell membrane.

The obtained Hi-C data [7] demonstrate that the 3D genome organizations of sperm and fibroblast cells show a high degree of similarity both with each other and with the previously described mouse embryonic stem cells. Topologically associated domains are present in spermatozoa and fibroblasts. Nevertheless, sperm cells and fibroblasts exhibit statistically significant differences between each other in the contact probabilities of defined loci. Tight packaging of the sperm genome results in an enrichment of long-range contacts compared with the fibroblasts. However, only 30% of the differences in the number of contacts are based on differences in the densities of their genome packages; the main source of the differences is the gain or loss of contacts that are specific for defined genome regions.

Conclusion: We considered a model the location of genes relative chromosome loops and binding sites. About 10% of genes have a beginning or an end in the same coordinates as the binding site and less than 2% for their outside. Genes of RefSeq are located inside the loop between the sites accounted for half of the total. While on a randomly generated set of binding sites, the figure was 30%.

Acknowledgements: The research has been supported by RFBR (№14-04-01906).

References:

1. E.V. Ignatieva et al. (2015) Regulatory Genomics: Combined Experimental and Computational Approaches, *Russian Journal of Genetics*, V. 51, - N. 4, P. 334–352.
2. G. Li et al. (2012) Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation, *Cell* V.148. – N.1-2. – P.84-98.
3. M.J. Fullwood et al. (2009) An oestrogen-receptor-alpha-bound human chromatin interactome, *Nature*. V. 462. – N. 7269. – P. 58-64.
4. J.R. Dixon, and B. Ren (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions, *Nature*, V. 485, P. 376–380.
5. E. Lieberman-Aiden et al. (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome, *Science*, V. 326, P. 289–293.
6. R. Kalhor et al. (2011) Genome architectures revealed by tethered chromosome conformation capture and population based modeling, *Nat. Biotechnol.,* V. 30, P. 90–98.
7. N. Battulin et al. (2015) Comparison of the 3D organization of sperm and fibroblast genomes using the Hi-C approach. *Genome Biol.* 16(1):77.

RECONSTRUCTION OF THE GENETIC STRUCTURE OF THE PROPOSED ANCESTRAL POPULATION OF TURKIC AND SOUTH SIBERIAN POPULATIONS

O.Yu. Kupova*[1], B.B. Yunusbayev[1,2]
[1]*Institute of Biochemistry and Genetics Ufa Scientific Center of the Russian Academy of Science, Ufa, Russia*
[2]*Estonian Biocentre, Tartu, Estonia*
e-mail: olgakupova@gmail.com
**Corresponding author*

Key words: Population genetics, Gene flow, Asia, Siberia, Turks

Motivation and Aim: According to the literature, almost all studied to date Turkic populations have a common pool of ancestors with populations of South Siberia and Mongolia (SSM), suggesting that the historical expansion of Turkic peoples occurred from this SSM area [1]. Despite of the clear signal of genetic contribution from the SSM area, it is difficult to estimate accurately the magnitude of the gene flow event. It is not clear whether the SSM ancestral population was admixed or was non-admixed Siberian population.

Methods and Algorithms: In this study, we analyzed published dataset of Illumina (650k SNPs) genotypes on various Turkic-speaking populations: the Anatolian Turks, Azeris of Iran and Dagestan, Balkars, Kumyks, Nogais, Gagauzes, Tatars, Bashkirs, Chuvashes, Kazakhs, Kyrgyz, Uzbeks, Turkmens, Karakalpaks, Uighurs, Yakuts, Dolgans; and populations of southern Siberia and Mongolia: Buryats, Tuvins, Altais, Khakassians, the Mongols of the northern provinces of China and Mongolia.

We conducted detection of IBD (identical by descent) chromosomal tracts (longer than 1cM) shared between South Siberian and Turkic populations using the Beagle 4.0 software [2]. We then carried out local ancestry inference for each loci within each detected IBD-segment using the PCAdmix software [3]. For our local ancestry inference, we used non-admixed Western Eurasian and East Eurasian (for example, the French and Chinese Han or Lithuanians and Nganasans) populations as surrogate ancestors.

Results: Our results showed that part of the IBD segments (~35%) represents a mosaic of haplotypes having West Eurasian and East Eurasian ancestry. We also show that haplotypes of West Eurasian ancestry constitutes of about 25% of total IBD tract length.

Conclusion: Assuming that IBD tracts represent a random sample of non-recombined chromosomal tracts descended from shared ancestors, we suggest a mixed origin of South Siberian ancestors for modern Turkic peoples.

Availability: Beagle 4.0 software is available for free at http://faculty.washington.edu/browning/beagle/beagle.html. PCAdmix software is available for free at https://sites.google.com/site/pcadmix/home.

References:

1. Yunusbayev B, Metspalu M, Metspalu E, Valeev A, Litvinov S. (2015) The Genetic Legacy of the Expansion of Turkic-Speaking Nomads across Eurasia. PLoS Genet 11(4):e1005068.
2. Browning BL, Browning SR. A fast, powerful method for detecting identity by descent. Am J Hum Genet. 2011;88(2):173–82.
3. Brisbin, A. (2010). Linkage Analysis for Categorical Traits and Ancestry Assignment in Admixed Individuals. Doctoral Dissertation, Cornell University, Ithaca, New York.

GENOME FRAGMENTS OF THREE APPLE VIRUSES ARE UNDER STRONG PURIFYING SELECTION

P.V. Kuzmitskaya*, O.Yu. Urbanovich
Institute of Genetics and Cytology of NAS, Minsk, Belarus
e-mail: P.Kuzmitskaya@igc.by
**Corresponding author*

Key words: apple chlorotic leaf spot virus, apple stem grooving virus, apple stem pitting virus, selection pressure, potential recombination events.

Motivation and Aim: Apple is one of the most widely grown and economically important fruit crop in Belarus. Apple mosaic virus (ApMV), apple chlorotic leaf spot virus (ACLSV), apple stem grooving virus (ASGV) and apple stem pitting virus (ASPV) are common in many cultivars. Often infected plants remain symptomless. However, apple viruses can cause significant yield reduction, up to 60%. These pathogens are single-stranded RNA viruses, and have high mutation rates compared with DNA organisms. The aim of this work was to estimate the genetic diversity of viral genomes fragments, evaluate selection pressure and detect and potential recombination events in these sequences.

Methods and Algorithms: RNA isolation from plant material was performed using GeneJetTM Plant Genomic RNA Purification Mini Kit (Thermo scientific (EC)) according to the recommended protocol. The fragments of the apple leafs were used for RNA extraction. The synthesis of the minus-strand cDNA was performed using RevertAidTM H Minus First Srtand cDNA Synthesis Kit (Thermo scientific (EC)) according to the manufacturer's protocol. Markers described in [1] were used for ASPV and ASGV amplification. To study the genetic diversity of the ACLSV primers F5 and R5 described in [2] were used. Fragments of the viral genomes were ligated into plasmid pTZ57R/T, and then E.coli DH5α was transformed with this plasmid. After cultivation on LB-Amp medium, the plasmid DNA was isolated from E.coli with Plasmid GeneJetTM Miniprep Kit (Thermo scientific (EC)). The viral fragments were sequenced using primers M13F and M13 R and BigDye Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems). CLUSTALW algorithm was used for multiple sequence alignment. All other analyses were performed with Datamonkey which is a front-end to HyPhy [3]. HKY85 model was used in this work.

Results: 5 isolates with nucleotide identities of 84,2% - 98,6% in fragments with overlapping ORFs coding replicase and movement protein were obtained from 5 trees infected with ACLSV. The length of each sequence is 850 bp. In the fragments of replicase gene the algorithm GARD identified potential recombination sites in the positions 237 and 405 of the sequences ($p = 0,05$), the algorithm SBP – in position 646. To determine the selective pressure on the certain sites of the sequences the phylogenetic trees constructed by both methods were used. Both methods SLAC and FEL didn't reveal the codons under positive selection. The number of sites under negative selection varied in the calculation of different methods from 167 to 182. Searching for signals of positive selection with the algorithm PARRIS confirmed their absence at the $p < 0.1$. According to GABranch results, the evolution of the studied sequences occurs at $dN / dS < 1$. It allow to suggest, that purifying selection dominates in the case of ACLSV replicase.

In the fragment of the ORF encoding the movement protein the potential recombination site was found. Algorithm GARD defined it in the position 67, while SBP– in 87. Evaluation of selective pressure identified that these nucleotide sequences are not under positive selection. The number of sites under negative selection varied

from 32 to 38 during calculation by different methods. PARRIS algorithm didn't show signals of positive selection at p <0.1. GABranch analysis showed that all branches of Neighbor-Jointing tree evolved at a value of dN/ dS = 0,034. The low values of dN/dS suggest that purifying selection occurred during analyzed history of ACLSV movement protein.

12 fragments of viral genome encoding the coat protein fragment and subsequent 3'-untranslated region were obtained from 9 individual tree infected with ASPV. The length of the obtained nucleotide sequences was 365- 367 bp. Identity of obtained sequences varying between 82 - 99.7%. The algorithm SBP identified the potential recombination site at position 55 in studied fragments, while the algorithm GARD didn't reveal such loci. Estimation of selection pressure with algorithms FEL and SLAC showed a lot of nucleotides under negative selection (77-71 depending on used method). There were no nucleotides under positive selection. The analysis results were confirmed by PARRIS analysis, which didn't reveal any evidence of positive selection at p <0.1 in the tested sequences. Grouping the branches of the evolutionary tree built by Neighbor-Joining method according to dN/dS values by GABranch algorithm showed that all branches evolved with value 0.011. It indicates that the evolution of the studied sequences was under strong purifying selection.

19 isolates with nucleotide identities of 96-100% in the fragment of coat protein gene were obtained from 11 trees infected with ASGV. The length of the obtained sequences was 272 - 275 bp. These fragments are located in C-terminal part of 241-kDa polyprotein and encode a viral coat protein. Searching recombination events in the sequence using the algorithm SBP identified possible site at the position 195, while the algorithm GARD didn't find the evidence of possible recombination processes. Evaluation of selective pressure conducted using algorithms FEL and SLAC didn't identify sites under positive selection. It was shown that the 24 - 43 nucleotides (depending on the used algorithm) in the fragments are under negative selection. PARRIS algorithm didn't show any signals of positive selection at p <0.1 in this dataset. GABranch analysis showed that all branches of Neighbor-Jointing tree evolved at the value of dN/ dS < 1. The low values of dN/dS suggest that purifying selection occurred during analyzed history of ASGV coat protein.

Conclusion: Used approaches shows that all studied fragment sequences of 3 apple viruses genomes are under strong purifying selection. No signals of positive selection were found in the studied sequences. Some potential recombination sites were identified in the studied fragments. Obtained results correspond to the results reported for other single-stranded RNA viruses of plants. Purifying selection is the main force of evolution of the members of the family Flexiviridae and studied fragments of ACLSV, ASPV and ASGV are no exceptions.

References

1. M. Hassan et al. (2006) Simultaneous detection and identification of four pome fruit viruses by one-tube pentaplex RT-PCR, Journal of Virological Methods 133: 124-129.
2. F. Niu et al. (2012) Complete nucleotide sequences of the genomes of two isolates of apple chlorotic leaf spot virus from peach (Prunus persica) in China, Archives of Virology, 157: 783-786.
3. S.L. Kosakovsky Pond, S.D.W. Frost (2005) Datamonkey: rapid detection of selection pressure on individual sites of codon alidnments, Bioinformatics, 21: 2531 - 2533 .

EXPRESSION ANALYSIS OF SOLOXOLONE METHYL EFFECTS ON CANCER CELLS

A.V. Markov[1*], E.B. Logashenko[1], A.E. Kel[2], O.V. Kel-Margoulis[2], M.A. Zenkova[1]
[1] *Institute of Chemical Biology and Fundamental Medicine SB RAS, Novosibirsk, Russia*
[2] *GeneXplain GmbH, Wolfenbuettel, Germany*
e-mail: markov_av@niboch.nsc.ru
**Corresponding author*

Motivation and Aim: Chemical transformation of natural bioactive compounds and metabolites is important area in drug development. Triterpenoids are considered as promising building blocks for synthesis of new highly active drugs due to their wide spectrum of biological activities and low toxicity [1]. New glycyrrhetinic acid derivative methyl 2-cyano-3,12-dioxo-11-deoxo-18βH-glycyrrhet-1(2),9(11)-dieneoate (Soloxolone methyl), was obtained by the direct modification of A- and C-rings of glycyrrhetinic acid. We showed that these modifications converted well-known triterpenoid glycyrrhetinic acid exhibiting weak antitumor activity to the derivative, which displays high antiproliferative activity with respect to cancer cells. It was shown that Soloxolone methyl (SM) induces cell death by caspase-dependent intrinsic apoptosis pathway [2]. To find the molecular targets of SM we investigated its effect on gene expression profile.

Methods and Algorithms: Human carcinoma cells KB-3-1 were incubated in the presence of 1 μM Soloxolone methyl for 1, 2, 3, 4, 6 and 10 h and then total RNA was extracted using TRIzol ® Reagent. Purified RNA was subjected to reverse transcription; cDNA probes were hybridized on HumanHT-12 v4 BeadChip (Illunmina, USA). Obtained raw data was analyzed using geneXplain platform 3.0 (GeneXplain GmbH, Germany) (http://genexplain-platform.com/bioumlweb/#). Differentially expressed genes were detected by hypergeometric analysis. Selected up- and down-regulated genes were subjected to functional analysis in Gene Ontology database. Transcriptional factors and master regulators of changed genes were identified by use of TRANSFAC 2012.3 and TRANSPATH databases, integrated in geneXeplain platform, as described in [3, 4].

Results: The analysis showed that cell incubation with SM resulted in significant time-dependent changes of expression levels of 311 genes (fold change >2, $p<10^{-3}$). Gene ontology analysis showed that these genes were mostly related ($p<10^{-10}$) to the following categories: response to endoplasmic reticulum stress, cellular response to stress, response to oxidative stress, regulation of apoptotic process, etc. Analysis of promoter [-500 bp, +100 bp] regions of detected up- and down-regulated genes in TRANSFAC database revealed a set of transcription factors (for instance, SRF, USF1, USF2, ATF3, JUN), which probably respond to SM. Using TRANSPATH database, we identified as well a set of master regulators (as example, p22phox, p70S6K2, pkmyt1, cxcr4, LAMA5) which seems like activated by SM.

Conclusions: The obtained data confirm the ability of SM to induce of apoptosis of cancer cells as shown previously [2]. Incubation of cancer cells in the presence of SM leads to EPR stress, overexpression of p53-regulated proapoptotic genes, activation of Egr1/ATF3/NAG1 and KLF4/p21 signaling pathways. Intracellular targets of SM were identified.

Ackowledgement: This work was supported by RAS programs "Molecular and Cellular Biology" and "Basic Sciences to Medicine" (No. 10), RFBI No. 12-04-31254, RFBI No. 14-03-3008, Scholarship of the President of the Russian Federation SP-408.2012.4, Grant in Support of Scientific Schools SS-1350.2014.4

References:

1. P. Dzubak et al. (2006) Pharmacological activities of natural triterpenoids and their therapeutic implications, Nat. Prod. Rep., 23(3): 394-411.
2. E.B. Logashenko et al. (2011) Synthesis and pro-apoptotic activity of novel glycyrrhetinic acid derivatives, ChemBioChem, 12(5): 784-94.
3. A. Kel et al. (2008) ExPlain: finding upstream drug targets in disease gene regulatory networks, SAR QSAR Environ. Res., 19(5-6): 481-94.
4. S.C. Chiu et al. (2010) Differential functional genomic effects of anti-inflammatory phytocompounds on immune signaling, BMC Genomics, 11: 513.

HIGH PERFORMANCE SIMULATION OF EVOLUTIONARY-POPULATION PROCESSES IN BACTERIAL COMMUNITIES

Z.S. Mustafin*, S.A. Lashin
Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
e-mail: mustafinzs@bionet.nsc.ru
**Corresponding author*

Motivation and Aim: The simulation of the evolution of bacterial communities is an actual problem of modern computational biology. Haploid Evolutionary Constructor (HEC) is a software package which is used for the simulation of evolution of prokaryotic communities. HEC has several levels of biological organization [1-2]: simulations of metabolism, substrate transport, interactions with environment and other populations. Some features of HEC population interpretation can bring big deviations in accuracy of final results. Communities can lose some part of their members. It can also break real biological variety of simulated community. The simulation bacterial communities of extremely high genetic diversity can take dozens of hours. In this study, we solve the problem of accuracy deviations by creating new population storage model and we solve the computational complexity problem by creating a number of different high performance algorithms (for CPU and GPU, for distributed-memory and shared-memory systems).

Methods and Algorithms: We have developed a novel model for population storage in HEC and a new algorithm for mutations simulation. The model completely solves the problem of accuracy loss in the HEC, and shows real biological variety of the simulated community. New mutation algorithm opens new opportunities for bacterial simulations. The storage model and mutation algorithm have been verified on the model of bacterial community. In this model, the fitness of community members simulated with the model of coadaptive alleles (i.e. species with appropriated allelic combinations had better bonus for utilization of substrate from environment). The problem was to study the conditions in which middle fitness individuals would be able to evolve into the highest fitness individuals, and to estimate the relation between fitness landscape function and evolutionary rate. For minimization of the simulation time, we have developed 3 high performance versions of population size changing function – the most time-consuming stage in HEC simulation process. First version uses the OpenMP technology for shared-memory systems. Second uses the MPI technology for distributed-memory systems. Third version was developed with the use of CUDA for simulations with using GPU. The tests and benchmarks have been performed on the clusters of the Center for collective usage "Bioinformatics" of the SB RAS (http://bioinformatics.bionet.nsc.ru/), and the Siberian Super Computer Center (http://www2.sscc.ru/). For shared-memory version testing and comparison used PC with AMD Phenom II X6 series CPU and Nvidia GTX 570 GPU).

Results: We got 4 parameters, affecting ability of species to mutate into better fitness group and affecting final part of the best group size in all population size: fitness landscape, probability of mutation, part of mutated species and fitness multiplier. We have shown that obtainment of evolutionary success in coevolving gene networks is determined by difference in fitness of different allelic combinations and probability of mutation. The part of mutating species determines the ability of species to mutate into the best group and not to die in a "bad" group between "best" and "middle" groups. Fitness multiplier determines the rate of group size and final part of group in all population. We have also shown that different fitness landscapes propose different optimal frequencies of mutation appearance; and either gradual or saltation evolutionary regimes can be optimal for them.

High performance algorithms released in this study, have reduced the time of simulation from some hours to some minutes. MPI and Open MP versions had near-linear acceleration (for MPI version it was also accepted with using 256 cluster threads). CUDA version showed a great decreasing of time of simulation of high-variety bacterial systems (time decreased from 200 second for one generation to 3 second). We created classification for best using of different high performance versions with different bacterial variety.

Conclusion: This work expands the HEC functionality. It allows to simulate different biological situations which can not be correctly simulated in the original HEC and led to the disturbance of accuracy and, sometimes, correctness of results. High performance algorithms allow user to simulate situations which are very difficult to simulate in vitro because of huge bacterial community variety and size. With new algorithms user can increase biological variety of simulating system without spending the additional time. New algorithms support simulations with using CPU, GPU, PC and high performance clusters.

Availability: http://evol-constructor.bionet.nsc.ru/?lang=en

Acknowledgements: The study was supported by the RFBR 15-07-03879 grant.

References:

1. Lashin S.A., Suslov V.V., Kolchanov N.A., Matushkin Yu.G. Simulation of coevolution in community by using the "Evolutionary Constructor" program. In Silico Biology, 2007, V 7, N 3, 261-275.
2. Lashin S.A. et al. Computer modeling of genome complexity variation trends in prokaryotic communities under varying habitat conditions, *Ecol. modelling,* **224:** 124-129., 2012

REPEATOME ANALYSIS OF SIBERIAN PINE

S.V. Novikova
Institute of Fundamental Biology and Biotechnology, Siberian Federal University, Krasnoyarsk, Russia
e-mail: kempfana@mail.ru

Key words: repetitive elements, genomic library construction, conifer genome, Siberian pine.

Motivation: Conifers have very large and complex genomes (12-30 Gb) containing up to 82 % of repeated sequences. Repetitive elements are extremely important for genome function. They are necessary in gene expression process of unique coding sequences and to organize additional functions essential for genome replecation. Repetitive DNA sequence elements are also fundamental to the cooperative molecular interactions forming nucleoprotein complexes. Deep investigation of Siberian pine repetitive elements yields more accurate genome structure and may be used for phylogenetic analysis and evolutionary studies.

Aim: The aim of the study is to find out repetitive elements, classify them and compare the data with the closely related species.

Methods and Algorithms: In this study, we used data from the whole genome sequencing of Siberian pine obtained in Laboratory of Forest Genomics at the Genome Research and Education Center in SibFU. We selected contigs exceeding 500 bp to make the operation easier and reduce time of programs run. We used Repeat Scout, Repeat Modeler, RepeatMasker programs, REPET pipeline, based on *de novo* and library-based approaches. Parallel work with a wide variety of software targeted to find out repeats allowed to reveal a significant portion of repetitive sequences.

Results: 2,025 types of repeats were found in the studied genome. Most of them, namely 1,623 are unknown due to insufficient study of repetitive elements of coniferous. Long terminal repeat-retrotransposons, such as Gypsy and Copia superfamilies comprised the most abundant fraction of classified repetitive elements. Also, we compared the generated library of repeats with the data on Siberian larch. The results show a high level of similarity of repeat content.

Conclusion: Further study of new repeats and implementation of repeats libraries for conifers are important issues for understanding the structure of the genome of Siberian pine and conifers in general. The study is a part of "Genomic studies major boreal coniferous forest tree species and their most dangerous pathogens in the Russian Federation" project funded by the Government of the Russian Federation (contract № 14.Y26.31.0004).

miRNAS BINDING WITH mRNAS OF *ARABIDOPSIS THALIANA* C2H2 TRANSCRIPTION FACTOR GENES

I.V. Pinsky*, A.T. Ivashchenko
Al-Farabi Kazakh National University, Almaty, Kazakhstan
e-mail: ilya.pinskyi@mail.ru
**Corresponding author*

Key words: plant, C2H2 transcription factors, miRNAs, mRNAs, CDS

Motivation and Aim: Transcription factors (TF) play a key role in gene expression regulation. 2290 transcription factors of *Arabidopsis thaliana* are included in 64 families. The C2H2 family of TF genes plays a key role in the plant development. Expression of transcription factor genes is controlled by miRNAs binding with mRNAs of genes and blocking translation. So it was important to study which miRNAs bind to mRNAs of *A. thaliana* C2H2 transcription factor genes, which genes are targets for each miRNA family and calculate free energy of miRNA-mRNA interaction.

Methods and Algorithms: The binding sites of ath-miRNAs in mRNAs of *A. thaliana* genes was revealed using the program MirTarget, which defines the location of the miRNA binding sites, calculates the free energy hybridization (ΔG, kJ/mole) and the ratio $\Delta G/\Delta G_m$ (%), where ΔG_m equal to the free energy of miRNA binding with fully complementary nucleotide sequence. The binding sites of miRNAs selected by $\Delta G/\Delta G_m$ equal or above to 85%. The diagrams of conservation of oligopeptides which are encoded by the miRNA binding sites are obtained by the program WebLogo.

Results: At present 337 miRNAs and 99 C2H2 TF genes are known in the genome of *A. thaliana*. It was found that only 49 miRNAs bind with mRNAs of *A. thaliana* C2H2 genes and only 59 genes are controlled by miRNAs. ath-miR5021 has the largest number of target genes. It bound with mRNAs of 18 genes with value of score from 87 to 93% and the binding sites of ath-miR5021 are encoded homological oligopeptide **SSSSS**. ath-miR5658 bound with mRNAs of 16 genes with value of score equal 85-96%, the binding sites are encoded homological amino acid sequence **HHHHH**. ath-miR414 bound with mRNAs of ten genes with value of $\Delta G/\Delta G_m$ equal 86-92%, the binding sites are encoded homological oligopeptide **DDDDD**. ath-miR854a bound with mRNAs of six genes with value of $\Delta G/\Delta G_m$ equal 85-96% and the binding sites are encoded homological oligopeptide **PPPPP**. ath-miR5015b bound with mRNAs of five genes with values of $\Delta G/\Delta G_m$ equal 86-88%. In the genome of Arabidopsis some miRNAs bound with mRNAs of one or several genes. ath-miR3434 bound with mRNAs of three genes (AT1G26610, AT3G58070 and AT4G35280) with score equal 86-94%. ath-miR838 bound with mRNAs of three genes too (AT2G18490, AT4G27240 and AT5G56200) with value of score ranging 86-90%. ath-miR860 bound with mRNAs of AT1G03840, AT3G50700 and AT5G44160 genes with value of score ranging 85-94%. ath-miR5632, ath-miR5638a, ath-miR5640, ath-miR830 and ath-miR837 miRNAs have two gene targets each. The rest 37 miRNAs have only a one gene target each. All miRNA binding sites are located in the protein-coding part (CDS) of C2H2 TF family mRNAs.

Conclusion: As a result of study of *A. thaliana* C2H2 TF gene family there was found that: 1) some target genes have more than one miRNA binding site, beginnings of which are located successively through one-three nucleotides that increases probability of translational process control by miRNA binding; 2) not all C2H2 family genes are targets for miRNAs; 3) nucleotides of miRNA binding sites encode oligopeptides of C2H2 proteins in different reading frames. Results obtained show that ath-miRNAs can regulate expression the most of C2H2 genes and influence the plant productivity and resistance.

Availability: miRTarget (our own program)

ARCHAEL AND BACTERIAL DIVERSITY IN HOT SPRING ALLA, BURYATIA, RUSSIA

A.A. Radnagurueva[1]*, M.R. Kabilov[2], E.V. Lavrentieva[1], B.B. Namsaraev[1]
[1]*Institute of general and experimental biology SB RAS, Ulan-Ude, Russia*
[2]*Institute of chemical biology and fundamental medicine SB RAS, Novosibirsk, Russia*
e-mail: aryuna_rg@mail.ru
**Corresponding author*

Key words: microbial community, bacterial diversity, archael diversity, pyrosequencing, hot springs, Alla, the Baikal Rift Zone

Motivation and Aim: Microbial communities of thermal habitats are under intense investigation that does not lose its relevance in recent decades. It is connected with the role played of microorganisms in the functioning of these extreme habitats and the discovery of new species of thermophilic representatives [1].

The main aim of the present study is to explore the composition and phylogenetic diversity of community's microbial mat, bottom sediments and water in hot spring Alla (Buryatia, Russia).

Methods and Algoritms: We have investigated the light-mineralized nitrogen alkaline hot spring Alla located on the territory of Barguzinskaya valley at the northeast side of the Lake Baikal (Russia). Alla hot spring flows directly into the channel of the Alla River on the rocky shallow. Spring water represents the sodium hydrocarbonate-sulfate type water at temperature of 73 °C and pH 9.7. Microbial mat, bottom sediments and water sample were fixed using 70% ethanol before DNA extraction and stored at 4°C. Water parameters were analysed *in situ*.

Total DNA from the studied samples were isolated by enzymatic lysis. DNA was extracted by the method of Marmur [2], further phenol purification of DNA was carried out as described by Maniatis et al [3].

Bacterial communities of thermal springs were studied with pyrosequencing of SSU rDNA amplicons. PCR amplification was performed with the use of eubacterial primers 343F and 806R. Reagents for PCR (DMSO, PCR buffer, polymerase, nucleotide triphosphates) were products of Thermo Scientific, USA. 50 µl of PCR mix contained 5ЧPhusion GC Buffer, 2.3 µM $MgCl_2$, 0.2 µM of each dNTP, 0.2 µM of forward and reverse primers, 50 ng of DNA and 0.7 u of Phusion High-Fidelity DNA Polymerase (Thermo Scientific, USA). The following amplification profile was used: 1 min at 98°C; 35 cycles of 15 s at 98°C, 15 s at 55°C, and 30 s at 72°C; and elongation phase of 5 min at 72°C. Amplified products were purified using MinElute Gel Extraction Kit (Qiagen, Netherlands). The resulted PCR fragments were purified by gel-electrophoresis in 1% agarose.

Pyrosequencing of DNA samples in the variable regions V1-V3 of the 16S rRNA gene was performed on a platform MiSeq (Illumina).

Obtained paired-end sequencing were bioinformatic processing with CLC GW 7.0 and usearch 7.0. Classification at phyla, classes, families and genera carried out using rdp classifier 2.10.1.

Results: The temperature and pH of the water in the sampling locations were 34-68°C and 9.3-9.9, respectively. The spring water for macrocomponental composition relates to HCO_3-SO_4-Na. The total mineralization of 0.2-0.26 mg/l.

In the community of microbial mat, bottom sediments and water in the hot spring Alla sequencing 16SrRNA received from 5295 to 41581 nucleotide sequences. The following eubacterial phyla are dominant: *Proteobacteria, Cyanobacteria, Acetothermia, Firmicutes, Chloroflexi, Deinococcus-Thermus, Nitrospirae*. The following representatives can be referred to the minor phyla: *Spirochaetes, Actinobacteria, Verrucomicrobia, Aquificae, Acidobacteria, Armatimonadetes, Planctomycetes* и *BRC1*. Among the archaea were found representatives types *Crenarchaeota* (0,1-4,7%).

The phylogenetic analysis has allowed classify absolute majority of microbial communities hot spring Alla at genus or family. Although phylogenetic proximity does not always correlate with the similarity of the metabolic pathways of microorganisms, these data allow us to propose hypotheses about the nature of the ecological relationships between the major groups of microorganisms.

The largest share in the community of microbial mats hot spring were representatives phyla *Atribacteria, Chloroflexi, Nitrospirae, Proteobacteria, Acidobacteria, Atribacteria* but their ratio varies significantly in different stations.

In the community of bottom sediments are generally dominated by representatives of phyla *Proteobacteria, Acetothermia, Atribacteria, Chloroflexi, Cyanobacteria, Aminicenantes, Acidobacteria, Atribacteria*. Comparative analysis of the dominant taxa in the bottom sediments of the investigated stations showed significant differences according to the temperature. It is shown that at a temperature of 64-68°C is dominated by the phylum *Firmicutes*, whereas when the temperature drops in the stations 2, 3 and 4 significantly reduced the proportion of *Firmicutes*. It also reduces the number of *Cyanobacteria* and increases *Chloroflexi*, that may indicate a change in the community when the temperature drops with cyanobacterial on anoxic.

In water is dominated *Deinococcus-Thermus, Proteobacteria, Cyanobacteria*. Taxonomic analysis of communities microbial mats, bottom sediments and water in the hot spring Alla showed that reflects the diversity of the communities totality of functional groups that perform various metabolic functions.

Conclusion: For the first time described in detail by sequencing the taxonomic composition and proportion of the major groups of microorganisms in various samples of hot spring Alla. The determining factor influencing the species diversity in the studied samples, most likely, is the temperature and the concentration of organic matter.

Acknowledgements: This work was supported by the Ministry of Education and Science of the Russian Federation (Project No. 1990) and Russian Foundation for Basic Research (RFBR), grant №15-04-01275a, №15-44-04335-r_sibir_a.

References

1. Zavarzin G.A. (2004) The study of microbial diversity in the Winogradsky Institute of Microbiology, Microbiology, 73(5): 598-612;
2. Marmur J. (1961) A procedure for isolation of deoxyribonucleic acid from microorganism, J. Mol. Biol., 3: 208-218;
3. Maniatis E. et al. (1984) Methods of genetic engineering, In: Molecular cloning, E. Maniatis et al. (Eds.), 480p.

MSCANNER - A GENOME-WIDE MOTIF FINDING TOOL

D.E. Romanov
Academy of Biology and Biotechnology, Rostov-on-Don, Russia
e-mail: rdme@ya.ru

Key words: motif finding, genome-wide motif finding.

Motivation and Aim: There are many motif finding algorithms each with its strong points wherther speed or accuracy. But it is often needed not just find motif, but also compare its location with location of other elements of a sequence, like genes, introns or promoters. The using of various custom formats of data representation hampers such kind of the research.

To illustrate this issue, lets consider the following list of motif finding software:

- GLAM2SCAN - a tool for finding occurrences of a gapped motifs in a sequence database [1];
- MAST - a tool for searching biological sequence databases for sequences that contain an occurrence of each motif in a given set of motifs [2];
- miRanda - an algorithm for the detection of potential microRNA target sites in genomic sequences [3].

These programs allow to find any kind of motifs, whereas miRanda was initially created to detect microRNA target sites. The problem here is that every of these programs uses its own custom input data format and produces output in its own custom format. Despite of input and output format of these programs being human readable, it makes complicated an automatic data processing.

Methods and Algorithms: To address this issues the software package **mscanner** was inroduced. **mscanner** reads motif sequences and genomic DNA sequences from input files and writes search results into the output file. Motif sequences should be in FASTA format and genomic sequences should be either in GENBANK or FASTA format. Output file is always written in GENBANK format.

By now only wrapping for GLAM2SCAN is implemented. Algorithm of finding motifs is made according the following procedure (quotation from GLAM2 software package manual): «By alternately applying glam2 and glam2mask several times, it is possible to find the strongest, second-strongest, third-strongest, etc. motifs in a set of sequences».

Another problem is that none of aforementioned motif finding programs (GLAM2SCAN, MAST and miRanda) could perform a whole genome motif search because of running out of system resources, in particular, the memory. To circumvent this issue the «divide and conquer» algorithm was applied. Initial long sequence was splitted into numerous overlapping fragments with length small enough to perform the search. Every fragment was searched for the occurences of the motifs and finding results were assembled then in a single output file. Such technique allows to study sequences with virtually arbitrary length.

Results: In order to verify the correctness of the proposed software pipeline the sequence of the human genome was scanned for occurrences of motifs of Alu elements. The computational results showed high consistency with publicly available experimental results of fluorescence *in situ* hybridisation of human chromosomes with probes for Alu elements.

It is worth noting, that the using of the standart data formats allows the researcher to *easily* watch the results in genome browsers, or to parse output files with help of such bioinformatic software packages, as BioPython or BioJava.

Availability: Available on request from the authors.

References

1. M.C. Frith, N.F.W. Saunders, B. Kobe, T.L. Bailey (2008) Discovering sequence motifs with arbitrary insertions and deletions, PLoS Computational Biology, 4(5): e1000071.
2. T.L. Bailey, M. Gribskov (1998) Combining evidence using p-values: application to sequence homology searches, Bioinformatics, 14(1): 48-54.
3. A.J. Enright, B. John, U. Gaul, T. Tuschl, C. Sander, D.S. Marks, (2003) MicroRNA targets in Drosophila,

STUDYING THE COMPLEXITY OF CONTEXTS SURROUNDING SNP SITES IN MAMMALIAN GENOMES

N.S. Safronova[1,2]*, I.V. Medvedeva[1,2], I.I. Abnizova[3]
[1] *Novosibirsk State University, Novosibirsk, Russia*
[2] *The Institute of Cytology and Genetics, Novosibirsk, Russia*
[3] *Welcome Trust Sanger Centre, Cambridge, United Kingdom*
e-mail: taschasafronova@mail.ru
**Corresponding author*

Key words: SNP, text complexity, polymorphisms

Motivation and Aim: Continuous data growth of genomic sequencing data raises new bioinformatics and technical tasks for analysis of natural genome variation, including SNPs (single nucleotide polymorphisms). A contextual study of physico-chemical and linguistic characteristics of genome sequences containing SNPs is one of such challenges. The aim of our work was to study the sequences containing SNPs in human, mouse and rat genomes using improved estimates of DNA text complexity. The data for human SNPs was taken according to the "1000 genomes" project, base assembly: GST Jan.2009 (GRCh37/hg19) and Dec.2013 (GRCh38/hg38). The data for mouse SNPs – assembly: Dec.2011 (GRCm38/mm10), the data for rat SNPs – assembly: Nov.2004 Baylor (3.4/rn4).

Methods and Algorithms: The data were downloaded from an open Internet-source UCSC Genome Bioinformatics database (http://genome.ucsc.edu/index.html). To study the contextual features of sequences we used the following mathematical measures and algorithms: Shannon's entropy, linguistic complexity (a measure that shows the richness of the text by different length "dictionaries"), the method according to Lempel-Ziv algorithm (a measure of the text compression) and a measure of monomers variability (a measure that reflects the degree of homogeneity of the text). These measures were implemented in the software package "Complexity_New" partially based on the algorithms of the software complex Complexity [1]. We analyzed more than 2.7 million SNP containing sites. The averaged text complexity profiles were calculated by program in C++. Emerging problem of context analysis of SNP positions relates to functional location of SNP in gene structure, in exons, protein-coding regions and functional sites. Thus, we need resources containing data on projections of protein functional sites on eukaryotic genes such as SitExa database that contains information on functional site amino acid positions in the exon structure of encoding gene [2].

Results: Research of contextual characteristics of sequences containing SNPs in human, mouse and rat genomes showed the local reduction in the complexity of the sequence in the area of polymorphism (+\- 20 nt). For the data for human genome the detected effect was confirmed as for genome assembly-2009 as for assembly-2013. Statistical analysis showed a high saturation of the polymorphism region with poly-A and poly-T tracts. An association between the distribution of SNPs and the presence of microsatellites is supported by several lines of evidence. Microsatellites exhibit high levels of length polymorphism such that heterozygous individuals can be viewed as carrying microdeletions, potentially enhancing the local mutation rate. Changes associated with the unusual base-stacking of purine-pyrimidine repeats or other structural properties could also be responsible for the mutational biases observed in regions flanking microsatellites. Generally, neighboring-nucleotide effects of SNPs depends on structural

content-based classifications (exonic, intronic) and sequence context-based categories (A/C, A/G, A/T, C/G, C/T, and G/T substitutions). We took into account all the variants and confirmed general low complexity trend around SNP position in range of 3-5 nt.

Conclusion: As we found, changes in complexity profiles are related to basic properties of DNA texts. Overall, low complexity profiles keep more information extending just measures of mononucleotide patches. The irregularities of mutation hot-spots in genome have been shown earlier on a limited data. The molecular mechanism of the observed effect of lowering the text complexity on flanks of SNP genome position can be explained by the increased frequency of double-helix DNA breaks in flanking positions. Overall, contextual study of the characteristics of sequences containing SNPs and the development of criteria for assessing the quality of sequencing data is an important task in bioinformatics. As it was previously shown [3] the short mononucleotide repeats can lead to erroneous readings while sequencing, so pre-processing of NGS data, which will then be published in the database, is very important task, especially for search for decease-associated SNPs [4]. The results obtained in this study could be explained by the physico-chemical properties of the DNA double helix [5]. It gives the reason to consider methods of an estimation of complexity of the text as one of the tools of analysis for the presence of sequencing errors. Our research extend observation of context-dependent SNP for wider taxonomy of model organisms [6,7] to the data obtained for natural variation in individual genomes data already collected at web-available databases.

Availability: the software is available by request to the corresponding author.

Acknowledgements: The research has been supported by ICG SB RAS budget project VI.61.1.2 and RFBR 14-04-0190.

Computing was done at Siberian Supercomputer center SB RAS (SSCC) and Shared Facility Center "Bioinformatics" ICG SB RAS.

References:

1. Y.L. Orlov, R. Te Boekhorst, I.I. Abnizova (2006) Statistical measures of the structure of genomic sequences: entropy, complexity, and position information, *J Bioinform Comput Biol*., **4**, 523-36.
2. I. Medvedeva et al. (2012) SitEx: a computer system for analysis of projections of protein functional sites on eukaryotic genes. *Nucleic Acids Res.*, **40**(Database issue), D278-83.
3. I. Abnizova et al. (2012) Analysis of context-dependent errors for illumina sequencing, *J Bioinform Comput Biol*., **10**(2), 1241005.
4. D.N. Cooper et al. (2011) On the sequence-directed nature of human gene mutation: the role of genomic architecture and the local DNA sequence environment in mediating gene mutations underlying human inherited disease. *Hum Mutat*. **32**(10), 1075-99.
5. K.J. Siddle et al. (2011) Bases adjacent to mononucleotide repeats show an increased single nucleotide polymorphism frequency in the human genome, *Bioinformatics,* **27**(7), 895-8.
6. F. Zhang, and Z. Zhao Z. (2004) The influence of neighboring-nucleotide composition on single nucleotide polymorphisms (SNPs) in the mouse genome and its comparison with human SNPs. *Genomics*, **84**(5), 785-95.
7. H. Zhao et al. (2005) Neighboring-nucleotide effects on the mutation patterns of the rice genome. *Genomics Proteomics Bioinformatics,* **3**(3), 158-68.

NUCLEIC ACIDS OF MAMMALIAN EXOSOMES

S.E. Sedykh*, V.V. Prints, V.N. Buneva, G.A. Nevinsky
Novosibirsk State University, SB RAS Institute of Chemical Biology and Fundamental Medicine, Novosibirsk, Russia
e-mail: sedyh@niboch.nsc.ru
**Corresponding author*

Key words: milk, exosomes, human milk, nucleic acids, protein complexes, supramolecular complexes

Motivation and Aim. Previously we obtained the stable supramolecular protein complex from human milk and placenta [1, 2]. We have shown that this complex contain major milk proteins and enzymes, and at the same time nucleic acids, lipids and oligosaccharides. To purify supramolecular complex we used the same methods which are used to isolate exosomes from milk. Now we try to compare the components of protein complex and exosomes and especially nucleic acids.

Methods. To purify protein complexes and exosomes from preparations of milk we use centrifugation and ultracentrifugation, ultrafiltration, gel-filtration. To elucidate the protein components of stable complexes and exosomes we use wide range of chromatography, electrophoresis and mass-spectrometry analysis. We plan to use Agilent Bioanalyser and Illumina MiSeq to detect nucleic acids in exosomes and stable protein complexes.

Results. With the light scattering and gel filtration we have shown that the complexes of human milk and placenta are stable in the presence of high concentrations of NaCl and $MgCl_2$, chaotropic agents, acetonitrile, detergents but dissociated efficiently under the conditions that destroy strong immune complexes (2 M $MgCl_2$, 0.5 M NaCl, and 10 mM DTT or 8 M urea, 50 mM EDTA, 0.5 M NaCl). The relative content of the individual protein complexes varied from 6% to 25% of the total milk protein. According to SDS-PAGE and MALDI-TOF-MS data, human placenta complex contains many major glycosylated proteins with low and moderate molecular masses 4-79.3 kDa, as well as minor proteins; human milk complexes contained lactoferrin and α-lactalbumin as major proteins, whereas human milk albumin and β-casein were present in moderate or minor amounts, also IgGs and sIgAs were observed. All stable protein supramolecular complexes contained DNA and RNA and efficiently hydrolyzed plasmid supercoiled DNA. According to our preliminary data, human, goat, cow and horse milk contain stable high molecular mass supramolecular protein complexes, which contains proteins and nucleic acids, similar to those in exosomes.

Conclusion. Exosomes of mammalian milk may contain nucleic acids and proteins similar to those in stable high molecular mass protein complexes. Sequencing of these nucleic acids will help to understand the mechanisms of exosome assembly and secretion.

References.

1. S.E. Soboleva, et. al (2015) Very stable high molecular mass multiprotein complex with DNase and amylase activities in human milk, Journal of Molecular Recognition, 28: 20-34.
2. E.E. Burkova et. al (2014) Extremely stable soluble high molecular mass multi-protein complex with DNase activity in human placental tissue, PLoS One, 9: e111234.

HOTSPOTS OF GENE REARRANGEMENTS AND TRAFFICKING IN MITOCHONDRIAL GENOME

E.A. Sirotinina*, D.Yu. Sherbakov
Irkutsk State University, Irkutsk, Russia
Limnological Institute SB RAS, Irkutsk, Russia
e-mail: haleo.inc@gmail.com
**Corresponding author*

Key words: gene rearrangements, pseudogene, mitochondrial genome

Motivation and Aim: The main direction of gene traffic leads from mitochondria to nucleus so that after the commitment of transfer the gene products are transported into the mitochondrion to perform their function. The set of genes in mitochondrial genome as well as the gene order varies strongly between major eukaryotic lineages. The repertoire of mitochondrial genes in most of cases is poorer then the repertoire of biochemical functions required. Transfer mechanism includes several steps, but the exact mechanism of deletion of mitochondrial "originals" coupled with the nuclear copies activation remains unknown. Transfers follow two general patterns: those that lead to resurrection of transferred gene function and those turning the transferred genes into mitochondrial pseudogenes [1]. Here we attempt to describe some common elements of evolutionary pathway of the rearranging mitochondrial genome in invertebrates.

Methods and Algorithms: We compared gene annotations resulting Hidden Markov Model method implemented MITOS pipeline [2] to the annotations of the same mitochondrial genomes meeting all GenBank [3] requirements such as uninterrupted reading frame (not ultimately required by MITOS). GenomeView [4] was used to compare prediction of genes based on HMM and MiTFi tRNA [5]. The procedure was applied to the set of published invertebrate mitochondrial genomes. The transformations detected were mapped on the phylogeny of the same taxa inferred from the sequences of nuclear genes.

Results: In course of the study many cases of apparently non-functional copies of mitochondrial genes were detected in mitochondrial genomes. Sometimes these loci were found on the opposite strand relative to a functional gene position in other taxa. The accumulation of nonsense mutations is believed to suggest that the gene is on it's way to final extinction. We have designated them as pseudogenes.

Some of genes picked up by MITOS pipeline, especially tRNA genes, had e-value insufficient to believe that they were functional. MiTFi approach have found additional tRNA genes in every mitochondrial genome, mostly in inverted position.

Recombination in the sites bearing combination of short reverse and direct repeats are hypothesized to increase the probability of strand swaps during recombination and thus serve as the hotspots for gene order rearrangements in mitochondrial genomes.

Conclusion: The traffic mechanism of genes between mtDNA strand in invertebrates and its reduction and pseudogene generation in mitochondrial genome are discussed.

Although the frequency of mtDNA recombination events in animals is very low and in fact unknown, it is important for understanding the rules guiding organization and functioning of mitochondrial genome.

References:

1. K.L. Adams, J.D. Palmer. (2003) Evolution of mitochondrial gene content: gene loss and transfer to the nucleus. Molecular Phylogenetics and Evolution, 29 : 380–395.
2. M.Bernt et al. (2013) MITOS: improved de novo metazoan mitochondrial genome annotation, Mol.Phylogenet. Evol. 69 : 313–319.
3. D.A.Benson et al. (2013) GenBank, Nucleic Acids Res. 41 : 36–42.
4. T.Abeel et al (2011) GenomeView: a next-generation genome browser, Nucleic Acids Research. doi: 10.1093/nar/gkr995.
5. F.Juhling, et al. (2012) Improved systematic tRNA gene annotation allows new insights into the evolution of mitochondrial tRNA structures and into the mechanisms of mitochondrial genome rearrangements, Nucleic Acids Res. 40 : 2833–2845.

GENE EXPRESSION DATA PROCESSING USING MICROARRAYS AND RNA PROFILING IN LABORATORY ANIMALS

A.M. Spitsina[1,2*], V.N. Babenko[2]
[1]*Novosibirsk State University, Novosibirsk, Russia*
[2]*Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia*
e-mail: anastasia.spitsina@gmail.com
**Corresponding author*

Key words: gene expression, microarray, gene network, human genome

Motivation and Aim: In recent years various databases has accumulated a large body of experimental data obtained using DNA microarrays in human as well as in laboratory animals [1]. Such data have great importance for applied medical research as well as for fundamental statistical analysis of gene expression in mammalian organisms across tissues and cell types. Thus, development of high-throughput computer tools for gene expression analysis should meet fast growth of available gene expression data [1,2]. Technical complexity of the work with data in this form consists in large volume and there are no a uniform, which make analysis difficult [1]. So the aim of this work was to develop and improve own program for data processing and analysis of gene expression.

Practical application of the software developed is in study of behavior in laboratory animals: mouse and rat [3]. To study genetic component of aggressive behavior we used published data on genes expression (RNA-seq and microarrays) related to aggressive behavior in mouse and rat as well own RNA-profiling data experimental data [3].

The research was intended to study the molecular and genetic mechanisms of enhanced aggressiveness in comparison with tolerant behavior using two unique experimental models which were developed at the Institute of Cytology and Genetics SB RAS: grey rats (*Rattus norvegicus*), which have been subjected to selection during several generations in two directions - friendly, tolerant behavior towards man, and increased aggressive behavior. The research aimed to elucidate the genetic and molecular mechanisms of hereditary defined (first model) and acquired (second model) increased aggressiveness using gene expression profiling by RNA-seq in different brain regions from aggressive and tame animals.

Methods and Algorithms: In this work we used Affymetrix expression microarray data for human, mouse *Mus musclus* and rat *Rattus norvegicus* to study gene co-expression patterns using data (suite consisting of two arrays) from BioGPS database. We developed C ++ software package which has options for statistical analysis and data pre-processing, such as the calculation and construction of tissue-specific profiles, filtering genes according to available information on the location on the chromosomes [2]. Also in our work we implemented a program for analysis of expression correlations (across samples and tissues), which will simplify the identification of structural features of gene networks. The program performs the analysis of the human genes expression, the study of the measurement signal quality on the microarray, analysis of tissue-specific gene expression, visualization of gene communication using correlation coefficients (linear Pearson and rank Spearman correlation coefficients) with a simple and intuitive interface. Visualization of the relationships between genes is constructed in the form of a gene network (using JavaScript). The program is applicable not only to the microarray data, but also to RNA-seq data (SRA (http://www.ncbi.nlm.nih.gov/sra), ArrayExpress (http://www.ebi.ac.uk/arrayexpress/).

Results: By developed software a comparative analysis of expression data for human genes whose expression is increased in the brain tissue was done. The expression patterns of pairs of transcripts, co-localized in the genome were analyzed [4]. Using RefSeq and BioGPS databases genes with high expression were identified, gene networks of interactions of these genes were constructed, correlation matrix and tissue specificity profiles of sample data were calculated. The structural features of genes with high expression were identified. After constructing differentially expressed gene lists in rat we used set of tools for coexpression analysis to found characteristic features of gene network related to aggressive behavior.

Conclusion: We studied correlation of gene expression contained in gene networks of circadian rhythm and cholesterol regulation, and well as genes responsible for aggressive behavior in mice (by previously annotated data). Also we made comparative analysis of the obtained results using program and information about co-expression of genes in the STRING database (http://string-db.org/). The effectiveness of gene networks reconstruction from the examined samples was analyzed. We confirmed many annotated gene network structures for curated gene sets.

RNA-seq analysis of differentially expressed genes in rat and mice confirmed presence of genes known as related to aggressive behavior, such as MaoA. We continue work on gene network reconstruction using RNA-seq experiments on additional mouse brain structures in contrast groups of laboratory animals [5].

Availability: https://github.com/amspitsina/gene-expr

Acknowledgements: The work was supported by RSF grant 14-14-00269.

References:

1. A.M. Spitsina, V.M. Efimov, V.N. Babenko, Y.L. Orlov (2014) Computer analysis of human gene expression data using BioGPS database of microarray Affymetrix U133 // In: Proceedings of the Ninth International Conference BGRS\SB-2014. Novosibirsk, Russia, June 23–28, 2014. Publishing House SB RAS. P. 154.
2. A.M. Spitsina et al. (2015) Supercomputer analysis of genomics and transcriptomics data revealed by high-throughput DNA sequencing, *Program systems: theory and applications* (http://psta.psiras.ru/) (In Russian) (*In press*).
3. N.N. Kudryavtseva, A.L. Markel, Y.L. Orlov (2014) Aggressive behavior: genetic and physiological mechanisms, *Vavilov journal of genetics and breeding*, 18(4/3):1133–1155. (In Russian)
4. Y.L. Orlov et al. (2012) ICGenomics: Program complex for analysis of symbol genomic sequences, *Vavilov Journal of Genetics and Breeding,* 16:732–741. (In Russian).
5. I.V. Medvedeva et al. (2014) Genome organization and context features of genes with over-expression in brain cells // Proceedings of XVI International Conference "Neuroinformatics-2014", Part 2, 32-42. MEPhI, Moscow. (In Russian)

COMPUTER ANALYSIS OF CLUSTERS OF TRANSCRIPTION FACTOR BINDING SITES IN MOUSE GENOME

A.V. Svichkarev*, A.I. Dergilev
Novosibirsk State University, Novosibirsk, Russia
The Institute of Cytology and Genetics, Novosibirsk, Russia
e-mail: svichkarev.anatoly@gmail.com
*Corresponding author

Key words: binding site, transcription factor, ChIP-seq, PFM, genome, DNA sequences

Motivation and Aim: Problem of coordinated gene expression regulation related to analysis of transcription factor binding sites location in human genome is important for wide range of biomedical applications [1]. Goal of this study is computer analysis of the nucleotide sequences and arrangement of binding sites for transcription factors obtained from the ChIP-seq experimental data. Several technologies have been developed last years to study the binding of transcription factors (TF) for transcription regulation, such as chromatin immunoprecipitation (ChIP) methods (ChIP-on-chip, ChIP-PET and ChIP-Seq). We used published data for location of binding sites of transcription factors in mouse genome (Chen et al., 2008) as well as in human genome available from GEO NCBI information resources. The question is to detect co-operative gene regulation by different transcription factors based on patterns of their binding sites co-localization in the genome. Such work is important to study the regulation of gene transcription in genome scale (research on embryonic stem cells, the expression of oncogenes). First step is detection of exact location of binding sites in nucleotide sequences using DNA binding motifs available from the databases (such as TRANSFAC, TRRD, JASPAR) [1].

Methods and Algorithms: In this work we used data on clusters of binding sites of different transcription factors [2]. Each cluster of coordinates in the genome contains from 4 to 11 different sites obtained by ChIP-seq experiment performed on mouse embryonic stem cells for 13 factors related to the pluripotency maintenance and cell differentiation. Peaks were identified earlier with limited accuracy using ChIP-seq profiles (about 100-200 nt) for mm8 genome release. We used published coordinates as well as recalculated clusters of binding sites for all the factors by own program. We used data on CTCF binding from [3]. To solve the problem of nucleotide sequence analysis independently developed software system in the language of Python 2.7 (IDE PyCharm 4.0.4). We used open source library and MOODS BioPython. MOODS is motif search algorithm in the nucleotide sequence. Our tools for the positional frequency matrix (PFM) analysis are implemented in C ++. BioPython provides a convenient means for data preprocessing and containers for the standard data formats. The software system is a set of modules. Each module is a command line tool with the help and supervision of input and output arguments. Consistent application modules allow to automate various stages of analysis and create different scenarios for integration with other services.

Results: In the course of this work was created and tested software localization motifs in the nucleotide sequence given in FASTA format, and in the format of genomic coordinates (in ChIP-seq peaks, BED files) and has additional features for format conversion and visualization of results. The project is under a free license GNU GPL v2.0.

Using the self-written software we searched exact localization of binding sites for 13 transcription factors (Oct4, Sox2, c-Myc, CTCF, E2f1, Esrrb, Klf4, Nanog, n-Myc, Smad1, STAT3, Tcfcp2i1, Zfx) by positional frequency matrix for each factor binding site detected by ChIP-seq in wide intervals (up to 500 nt) in the mouse genome. Experimentally determined numbers of sites range from a few thousand to tens of

thousands. The percentage of the presence of motifs in ChIP-seq peaks ranged from 25% to 99%. Using the own software we found some preferences in location of transcription factor binding sites of various types (Oct4, Nanog, Sox2) and TF n-Myc and c-Myc. We calculated the distance between the nearest group of TF binding sites and the distance to nearest gene promoters.

Conclusion: Exact localization of binding sites for 13 transcription factors (Oct4, Sox2, and others) were revealed based on ChIP-seq data and positional frequency matrix for each factor binding sites. We created and tested software for localization of motifs inside ChIP-seq peak (genome coordinates). Some binding motifs were iteratively refined using previously known PFM. Binding sites of some TFs tend to be in promoter regions of genes. All the individual sites and clusters were separated on groups populating proximal promoter and distal regions relative to target (nearest) gene. We counted about 20% binding sites in gene promoter regions and ~80% sites residing inside gene borders and in distal regions. The presence of preferences in position transcription factor binding sites in clusters of different types of transcription factor was shown. Location of transcription factor binding sites in distal gene regions could be further investigated by modern technologies reveling distal regulatory chromosome contacts such as ChIA-PET [4].

Availability: https://github.com/SvichkarevAnatoly/Bioinformatics-DNA-Motifs-Search

Acknowledgements: The research has been supported by RFBR (№14-04-01906) and ICG SB RAS budget project VI.61.1.2. Computing was done at Siberian Supercomputer center SB RAS (SSCC) and Shared Facility Center "Bioinformatics" ICG SB RAS.

References:

1. E.V. Ignatieva et al. (2015) Regulatory Genomics: Combined Experimental and Computational Approaches, Russian Journal of Genetics, V. **51**, N. 4, 334–352.
2. X. Chen et al. (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells, *Cell*, **133**(6), 1106-17.
3. G. Li et al. (2014) Chromatin Interaction Analysis with Paired-End Tag (ChIA-PET) sequencing technology and application, *BMC Genomics,* **15**(Suppl 12), S11.
4. L. Handoko et al. (2011) CTCF-mediated functional chromatin interactome in pluripotent cells, *Nat Genet*, **43**(7), 630-638.

IDENTIFICATION OF FUNCTIONAL NETWORKS ASSOCIATED WITH CELL DEATH IN THE RETINA OF OXYS RATS DURING THE DEVELOPMENT OF RETINOPHATHY

D.V. Telegina*, E.E. Korbolina, N.I. Ershov, O.S. Kozhevnikova
Institute of Cytology and Genetics, Novosibirsk, Russia
e-mail: daria.victorovna@gmail.com
**Corresponding author*

Key words: age-related macular degeneration, retinopathy, apoptosis, RNA-seq, OXYS rats

Purpose: Age-related macular degeneration (AMD) is a major cause of blindness in developed countries, and the molecular pathogenesis of early events in AMD is poorly understood. Senescence-accelerated OXYS rats develop AMD-like retinopathy. The aim of this study was to explore the differences in retinal gene expression between OXYS and Wistar (control) rats at age 20 days and to identify the pathways of retinal cell death involved in the OXYS retinopathy initiation and progression.

Methods: Retinal mRNA profiles of 20-day-old OXYS and Wistar rats were generated at the sequencing read depth 40 mln, in triplicate, using Illumina GAIIx. A terminal deoxynucleotidyl transferase-mediated deoxyuridine triphosphate nick end labeling (TUNEL) assay was performed to measure the apoptosis level. GeneMANIA was used to construct interaction networks for differentially expressed (DE) apoptosis-related genes at ages 20 days and three and 18 months.

Results: Functional analysis was suggestive of a developmental process, signal transduction, and cell differentiation as the most enriched biological processes among 245 DE genes at age 20 days. An increased level of apoptosis was observed in OXYS rats at age 20 days but not at advanced stages. We identified functional clusters in the constructed interaction networks and possible hub genes (*Rasa1*, *cFLAR*, *Birc3*, *Cdk1*, *Hspa1b*, *Erbb3*, and *Ntf3)*. We also demonstrated the significance of the extrinsic apoptotic pathway at preclinical, early, and advanced stages of retinopathy development. Besides the cell death signaling pathways, immune system-related processes and lipid-metabolic processes showed overrepresentation in the clusters of all networks.

Conclusions: These characteristics of the expression profile of the genes functionally associated with apoptosis may contribute to the pathogenesis of AMD-like retinopathy in senescence-accelerated OXYS rats.

HIGH-PERFORMANCE COMPUTATIONS SUPPORT FOR THE SOFTWARE PACKAGE «HAPLOID EVOLUTIONARY CONSTRUCTOR»

R.K. Zudin*, S.A. Lashin
Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
Novosibirsk State University, Novosibirsk, Russia
e-mail: rockalocular@gmail.com
**Corresponding author*

Motivation and Aim: Modelling of population dynamics and evolutionary-genetic processes is an important problem of modern mathematical biology. To solve this problem, a software package «Haploid Evolutionary Constructor» has been developed in ICG SB RAS (hereinafter - HEC) [1]. HEC provides an opportunity to simulate the functioning and evolution of prokaryotic communities, considering the following factors: the influence of environment, trophic relationships between populations and speciation. The number of communities can be as high as 10^{20} cells.

Biological objects are complex, hierarchically organized systems and their models are complex correspondingly. At the same time, the interaction between these biological objects, taking into account the spatial distribution, combined with the large number of populations and individuals within them, promotes increasing complexity of a computational problem. As a result, calculations of such models require a large amount of time. Therefore, time required for calculations should be reduced without losing the accuracy and adequacy of simulation results.

The aim of this work is the organization of high-performance computing for the software package «Haploid Evolutionary Constructor». We have developed several high-performance versions for HEC. Parallelization efficiency was obtained up to 80% while acceleration was obtained up to 50 times.

Methods and Algorithms: Parallelization was realized using MPI and QTConcurrent technologies for computer cluster and desktop usage respectively.

Results: HEC was optimized using MPI and QTConcurrent technologies; Optimized versions of HEC were tested using real and massive data; Analysis of results, obtained during tests, which showed that the speed of the original program was increased up to 50 times in case of one dimensional redistribution of populations and substrates; and up to 10 times in case of two/three dimensional redistribution; Analysis of parallelization efficiency, which showed 7% efficiency using the maximum number of processes and 80% using minimum number of processes (in case of one dimensional redistribution of populations and substrates); 8% (maximum number of processes) and 76% (minimum number of processes) in case of two dimensional redistribution; 4% (maximum number of processes) and 55% (minimum number of processes) in case of three dimensional redistribution.

Conclusion: Thus, optimized version of HEC allows to reduce the time of calculations in comparison with the original program.

Acknowledgements: The study was supported by the RFBR 15-07-03879 grant.

References:

1. Lashin S.A., Matushkin Yu.G. Haploid evolutionary constructor: new features and further challenges // In Silico. Biol. 2012. V.11. N. 3. P. 125-135.

CTCF BINDING SITES AND GENE EXPRESSION ANALYSIS IN GENOME SCALE BY SEQUENCING DATA

E.V. Kulakova[1,2], A.M. Spitsina[1], N.S. Safronova[1], A.V. Svichkarev[1], A.I. Dergilev[1], M.P. Ponomarenko[2], V.N. Babenko[2], P. Hong[3], G. Li[3], Y.L. Orlov[1,2]*

[1]*Nosobirsk State University, Novosibirsk, Russia*

[2]*Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia*

[3]*Huazhong Agricultural University, Wuhan, China*

e-mail: orlov@bionet.nsc.ru

[*]*Corresponding author*

Key words: next-generation sequencing, transcription factor binding sites, chromosome contacts, genome, CTCF

Motivation and Aim: Identification of genome-wide distal chromatin interactions serves as background for transcription regulation studies. Hi-C technology allows the finding of chromatin interaction in genome scale, but needs higher resolution for detection of contacting sites. Chromatin Interaction Analysis with Paired-End-Tag sequencing (ChIA-PET) method [1] fits these demands, but requires the development of specialized high-throughput software for data integration and statistical estimation of the data. The study of chromosomal contacts raises the problem of statistical analysis, which requires the processing of huge volume of experimental data. The aim of the work was to develop a computer program for statistical data analysis and test it on CTCF (CCCTC-binding factor) binding sites data and topological domains [2].

Methods and Algorithms: We used data on the spatial domains in the genome of the mouse embryonic stem cells and in the human genome, data on the location of CTCF binding sites clusters obtained by ChIA-PET as well as obtained experimentally by ChIP-seq. The processing of Hi-C data includes the isolation of the spatial chromosomal domains; such information is stored in databases such as 3DGD (http://3dgd.biosino.org).

Results: We constructed the distribution of CTCF transcription factor binding sites on topological domains on the human chromosomes. The distributions of human genes relative CTCF binding sites and a randomly generated list of such sites as the program output were used to estimate statistical significance of the associations found.

Conclusion: We considered a model with the location of genes relative to chromosome loops and binding sites. High conservation of CTCF binding motifs and "Forward-Reverse" location of CTCF binding sites pairs in contacting chromosome regions was confirmed. We show that CTCF loops encircled gene bodies.

Acknowledgements: The research has been supported by RSF (№14-24-00123). Computing was done at Shared Facility Center "Bioinformatics", SSCC SB RAS.

References:

1. G. Li et al. (2012) Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation, *Cell,* V.148. – N.1-2. – P.84-98.
2. L. Handoko et al. (2011) CTCF-mediated functional chromatin interactome in pluripotent cells, *Nat Genet*, V. 43. – N.7. – P.630-638.

COMPUTER ANALYSIS OF NATURAL ANTISENSE TRANSCRIPTS IN EUKARYOTIC GENOMES

A.M. Spitsina[1,2], N.S. Safronova[1], E.V. Kulakova[1,2], A.I. Dergilev[1], Q.L. Li[3], J.J. Wang[3], D.J. Chen[3], C.H. Yuan[3], M.P. Ponomarenko[1,2], D.A. Afonnikov[2], Y.L. Orlov[1,2]*, M. Chen[3]

[1]*Nosobirsk State University, Novosibirsk, Russia*
[2]*Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia*
[3]*Zhejiang University, Hangzhou, China*
e-mail: orlov@bionet.nsc.ru
[*]*Corresponding author*

Key words: sequencing, transcription, plant genomics, natural antisense transcripts

Motivation and Aim: Computer integration of next generation sequencing data in specialized databases is important in system biology and bioinformatics. Pairs of RNA molecules transcribed from partially or entirely complementary loci are called cis-natural antisense transcripts (cis-NATs), and they play key roles in gene expression regulation. Natural antisense transcripts (NATs) are capable of regulating the expression of target genes at different levels (transcription, mRNA stability, translation). They have been identified in multiple eukaryotes, including humans, mice, yeast and plants.

Methods and Algorithms: The study presents computer analysis of plant NATs, including availability, conservation and functions. Earlier, identification of chromatin signature of cis-NATs in Arabidopsis allowed to suggest a connection between cis-NAT transcription and chromatin modification in plants. An analysis of small-RNA sequencing data showed that ~4% of cis-NAT pairs produce putative cis-NAT-induced siRNAs. We developed set of computer programs to analyze cis-NAT pairs and miRNA genes. Text complexity as a measure of context dependencies was applied for nucleotide sequences containing mapped cis-NATs in plants, as previously we used it for monomer repeats analysis.

Results: Presence of low complexity regions in genomes containing NATs was shown. We have analyzed data from PlantNATsDB (Plant Natural Antisense Transcripts DataBase) published platform for annotating and discovering NATs by integrating various data sources. The database provides an integrative, interactive web graphical interface to display multidimensional data, and facilitate research and the discovery of functional NATs. We compared gene structures containing NATs for wheat and related plant genomes.

Conclusion: A method for large-scale detection of sense-antisense transcript pairs involves microarrays data analysis (Affymetrix Wheat GeneChip). About hundred sense-antisense transcript pairs were found. Analysis of the gene ontology terms showed a significant over-representation of transcripts involved in energy production. Appearance of new sequencing technologies such as Hi-C and ChIA-PET extends analysis of overlapping transcripts to more complex models.

Acknowledgements: The work is supported in part by Sino-Russian RFBR grant 15-54-53091 and ICG SB RAS (Budget project VI.61.1.2). Computing was done at Shared Facility Center “Bioinformatics”, SSCC SB RAS.

Author Index

Научное издание

Тезисы 7-й Международной школы молодых ученых
«Системная биология и биоинформатика», SBB-2015

на английском языке

Abstracts of the 7th International Young Scientists School
“SYSTEMS BIOLOGY AND BIOINFORMATICS”

In the original wording

Подготовлено к печати
в редакционно-издательском секторе
ИЦиГ СО РАН
630090, Новосибирск, пр. акад. М.А. Лаврентьева, 10

Дизайн и компьютерная верстка: А.В. Харкевич
Подписано к печати 15.06.2015 г.
Формат бумаги 70 × 108 1/16. Печ. л. 4,9. Уч.-изд. л. 5,5
Тираж 100
Отпечатано на полиграфической базе ИЦиГ СО РАН