RUSSIAN ACADEMY OF SCIENCES
SIBERIAN BRANCH

INSTITUTE OF CYTOLOGY AND GENETICS

THE NINTH INTERNATIONAL CONFERENCE ON BIOINFORMATICS OF GENOME REGULATION AND STRUCTURE\SYSTEMS BIOLOGY

Abstracts

BGRS\SB-2014
Novosibirsk, Russia
June 23–28, 2014

Publishing House SB RAS

Novosibirsk
2014

INTERNATIONAL PROGRAM COMMITTEE

Nikolay Kolchanov, Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia (Chairman of the Conference)
Ralf Hofestädt, University of Bielefeld, Germany (Co-Chairman of the Conference)

Alexandr Ratushny, Seattle BioMed and Institute for Systems Biology, Seattle, USA
Andrey Rzhetsky, The University of Chicago, USA
David Liberles, University of Wyoming, USA
Dmitry Afonnikov, Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
Egor Prokhorchuk, Bioengineering Center RAS, Moscow, Russia
Elena Salina, Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
Evgenii Rogayev, University of Massachusetts, USA
Evgeny Nikolaev, Institute of Energy Problems of Chemical Physics RAS, Moscow, Russia
Frank Eisenhaber, Bioinformatics Institute, A-STAR, Singapore
Irina Larina, SSC RF Institute for Biomedical Problems, RAMS, Moscow, Russia
Konstantin Anokhin, NRC "Kurchatov Institute", Moscow, Russia
Konstantin Skryabin, Bioengineering Center RAS, Moscow, Russia
Lars Kaderali, Dresden University of Technology, Germany
Limsoon Wong, National University of Singapore, Singapore
Maria Samsonova, Saint Petersburg State Polytechnic University, Saint Petersburg, Russia
Maxim Fedorov, University of Strathclyde, UK
Ming Chen, Zhejiang University, China
Nikolay Podkolodnyy, Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
Vladimir Ivanisenko, Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
Vsevolod Makeev, Vavilov Institute of General Genetics RAS, Moscow, Russia
Yijun Ruan, The Jackson Laboratory, USA
Yuriy Orlov, Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
Yury Matushkin, Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia

LOCAL ORGANIZING COMMITTEE

Svetlana Zubova, ICG SB RAS (Chairwoman)
Andrey Kharkevich, ICG SB RAS
Anna Onchukova, ICG SB RAS
Erlan Tokpanov, ICG SB RAS
Galina Kiseleva, ICG SB RAS
Ilya Akberdin, ICG SB RAS
Nadejda Glebova, ICG SB RAS
Dmitry Oshchepkov, ICG SB RAS
Evgenia Oshchepkova, ICG SB RAS
Olga Popik, ICG SB RAS
Tatyana Karamysheva, ICG SB RAS

CONTACTS

Organizing committee: bgrs2014@bionet.nsc.ru
The Siberian Branch of the Russian Academy of Sciences
RUSSIAN FEDERATION
630090 Novosibirsk, Lavrentyeva, 10
Tel: +7(383) 363-49-80
Fax: +7(383) 333-12-78
URL - ICG SB RAS: www.bionet.nsc.ru
URL - BGRS\SB-2014: http://conf.nsc.ru/BGRSSB2014

ISBN 978-5-7692-1371-7 (T. 1)
978-5-7692-1370-0

Organizers

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

Chair of Information Biology, Novosibirsk State University (NSF NSU)

Siberian Branch of the Russian Academy of Sciences

Company «Scientific service» Ltd.

The Vavilov Society of Geneticists and Breeders

The program "Genomics, proteomics and bioinformatics" of the SB RAS

Sponsors

Russian Foundation for Basic Research

SILVER SPONSORS

Helicon

Limited Liability Company Eppendorf Russia

Genotek

Bio-Rad

Intel

InterLabService

BRONZE SPONSORS

Roshe diagnostics

CONTENTS

COMPUTER HIGH-THROUGHPUT APPROACHES TO WHEAT PHENOTYPING

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

Motivation and Aim: The growing need for rapid and accurate approaches for large-scale assessment of phenotypic characters in plants becomes more and more obvious in the studies looking into relationships between genotype and phenotype. This need is due to the advent of high throughput methods for analysis of genomes. Nowadays, any genetic experiment involves data on thousands and dozens of thousands of plants. Traditional ways of assessing most phenotypic characteristics (those with reliance on the eye, the touch, the ruler) are little effective on samples of such sizes. Modern approaches seek to take advantage of automated phenotyping, which warrants a much more rapid data acquisition, higher accuracy of the assessment of phenotypic features, measurement of new parameters of these features and exclusion of human subjectivity from the process. Additionally, automation allows measurement data to be rapidly loaded into computer databases, which reduces data processing time. In this work, we present the WheatPGE information system designed to solve the problem of integration of genotypic and phenotypic data and parameters of the environment, as well as to analyze the relationships between the genotype and phenotype in wheat.

Methods and Algorithms: The current data model is represented by 27 tables related to each other. MySQL is used as a server. To work with a database, we developed a webinterface that is implemented on the basis of the module Catalyst—a free cross-platform software framework written in Perl. To process images obtained from a camera the open library of computer vision Open CV was used. Android platform was chosen to develop the mobile applications.

Results and Conclusion: The system is used to consolidate miscellaneous data on a plant for storing and processing various morphological traits and genotypes of wheat plants as well as data on various environmental factors. The system is available at www.wheatdb.org. Its potential in genetic experiments has been demonstrated in high-throughput phenotyping of wheat leaf pubescence.

This work supported by Programs A.II.6 from the Russian Academy of Sciences, RFBR grant №14-07-31226, budget project VI.61.1.2.

KINETIC MODELING OF PYRIMIDINE BIOSYNTHESIS IS A FIRST STEP TO *IN SILICO* BACTERIAL CELL

Akberdin I.R.[1]*, Ermak T.V.[1], Kazantsev F.V.[1], Khlebodarova T.M. [1]
Likhoshvai V.A.[1,2]
[1]*Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia*
[2]*Novosibirsk State University, Novosibirsk, Russia*
e-mail: akberdin@bionet.nsc.ru
**Corresponding author*

Key words: *in silico* cell, kinetic modeling, *E.coli*, pyrimidine biosynthesis

Motivation and Aim: The idea of reconstruction a comprehensive *in silico* cell began to be realized after the organisms with extremely small genome size were found and models of the minimum cell including some models for more complex organisms were developed.

One of the key tool for its creation is a mathematical modeling, which allows us to naturally combine the experimental data on regularities of cell functioning on all levels of their organization in framework of unified conceptual scheme [1]. According to the scheme, a mathematical modeling of complex cell systems can be performed as a sequential process of their description through elementary subsystems. We have used the approach for modeling of pyrimidine biosynthesis in *E. coli* cell. The nucleotides metabolism is a fundamental cell subsystem, which provide a metabolic supply for replication and transcription processes.

Methods and Algorithms: Kinetic models of enzymatic reactions of the pyrimidine biosynthesis pathway were developed on basis of the mass-action law, Michaelis–Menten equation and generalized Hill functions approach [2]. Values of 88 model parameters were extracted from published data; values of three parameters were estimated on basis of the known analogue parameters for other bacteria. Analysis of the individual contribution of each negative feedback in the mode of pyrimidine biosynthesis functioning carried out by generation of the reduced versions of the model with different sets of regulatory negative feedbacks from 1 to 8.

Results: The mathematical model of pyrimidine biosynthesis in *E. coli* cell has been developed including description of known negative feedbacks that regulate dynamics of the system functioning. The mutational portrait of the model has been constructed using of that the limited enzymatic reactions have been revealed. The parametric domains of existence in the system of stationary and oscillatory modes as well as the impact of uridine-5-monophosphate catabolism on the dynamics of the pathway functioning in the *carB* and *ΔpyrI carB* mutations have been investigated [3].

Acknowledgements: The study was partially supported by RFBR grant [13-01-00344] and Presidium RAS Program “Molecular and cell biology” (6.8).

References:

1. Akberdin I.R. *et al.* (2013). *Mat.Biol.&Bioinf.(Russian)*, 8(1), 295-315.
2. Likhoshvai V.A., Ratushny A.V. (2007) *J. Bioinform. Comput. Biol.,* **5**: 521-531.
3. Reaves et al. (2013) *Nature.* doi:10.1038/nature12445

REGULATORY MECHANISMS FOR MESC SELF-RENEWAL: KINETIC AND STOCHASTIC MODELING

Akberdin I.R.*, Ivanisenko N.V., Oshchepkova E.A., Omelyanchuk N.A., Matushkin Yu.G., Afonnikov D.A., Kolchanov N.A.
Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
e-mail: akberdin@bionet.nsc.ru
**Corresponding author*

Key words: pluripotent state, differentiation, gene expression, NANOG heterogeneity, modeling

Motivation and Aim: The current experimental data supports the hypothesis that the Nanog-low state is a transient and reversible state that acts as a temporally restricted ‘‘gate-keeper’’ for extrinsic signals by which embryonic stem cells are directed towards differentiation Therefore, the molecular control of Nanog levels has to be regarded as an important piece in the puzzle of pluripotency organization as it is a potential candidate mechanism to maintain the balance between self-renewal and differentiation. A detailed quantitative understanding of these processes and their consequences on the system dynamics would be highly beneficial for the development and the optimization of protocols for generation, maintaining and differentiation of induced pluripotent and embryonic stem cells. Nevertheless, existing models describing core genetic network regulation and Nanog heterogeneity have a number of drawbacks, as they contain hardly biologically interpretable parameters and do not include explicit molecular mechanisms of Nanog regulation based on recent experimental data [1].

Methods and Algorithms: The structure of the model was taken from the work [2] with modification accounting for the Nanog autorepression [1]. The kinetic model is reconstructed using the generalized chemical-kinetic approach as well as a generalized Hill functions method [3]. Simulations of stochastic model’s variant were performed using the Next Step method of the Gillespie Algorithm, designed specifically for the simulation of coupled chemical reactions [4].

Results and Conclusion: The main factors required for the self-renewal of murine ESCs are externally applied BMP and LIF and internally expressed *Oct4*, *Sox2* and *Nanog* genes. Interactions between these primary participants are still poorly understood. We have analyzed via kinetic and stochastic modelling a several hypotheses on these interactions assumed from data published recently. Our preliminary simulation results allow us to suggest some factors influencing Nanog heterogeneity. The revealed fact is necessary to confirmed experimentally.

Acknowledgements: The work supported by Skolkovo Center for Stem Cell Research (Agreement with IOGen RAS #8418-43/2013).

References:

1. Navarro P. *et al*. (2012) *EMBO J*, **31**: 4547–62;
2. Jaenisch R., Young R. (2008) *Cell*, **132**(4): 567-582;
3. Likhoshvai V., Ratushny A. (2007) *JBCB*, **5**(02b): 521-531;
4. Gillespie D. T. (1977) *JPC,* **81**(25): 2340-2361.

HIDDEN RESERVES OF USED VACCINE SUBSTRAIN

Aksenova E.I.*, Voronina O.L., Kunda M.S., Semenov A.N., Zamyatnin A.A., Lunin V.G., Gintsburg A.L.
Gamaleya Scientific Research Institute for Epidemiology and Microbiology Ministry of Health of Russian Federation, Moscow, Russia
e-mail: eaksenova@list.ru
**Corresponding author*

Key words: Vaccine substrains *Mycobacterium bovis* BCG Russia (BCG-Russia) and *M. bovis* BCG Tokyo (BCG-Tokyo), bioinformatic analysis, hypothetical proteins

Motivation and Aim. Comparison of sequenced genomes vaccine substrains BCG-Russia and BCG-Tokyo and detection of potential of hypothetical proteins.

Methods and Algorithms. Whole genome sequencing (WGS) of BCG-Russia was made in our institute on the 454 GS Junior (Roche) via two methods: shotgun and pair end sequencing. Annotation was carried out in RAST (Rapid Annotation using Subsystem Technology) service, applicable for GC-rich genomes. Mapping of genome was made using Whole Genome Mapping technology (OpGen). Analysis of fragmentome *in silico* was conducted by comparing of primary structures of protein fragments with amino acid sequences of natural regulatory oligopeptides presented by the EROP_Moscow (Endogenous Regulatory OligoPeptide knowledgebase) database.

Results. According to data of WGS, the coverage of BCG-Russia genome was 67,3 and over 99% of genome was assembled in 4 extended scaffolds. The similarity with genome of closely related substrain BCG-Tokyo was more than 98% including large duplication region (DU2) – 3 copies as in reference. Differences were minimal: 3 deletions (1-3 bp), 7 insertions (1-22 bp), 49 SNPs.

According to RAST annotation, in the BCG-Russia genome there are 4381 open reading frames. 1119 ORFs encodes hypothetical proteins (37-855 aa), among which 86% had analogues in database. Analysis via servers KEGG, BLAST, COGs, allowed to detect conserved domains in homologues' structure for more than a half of hypothetical proteins: DUF (domain of unknown function), enzymes, toxin/antitoxin, transcription factors et al. Signal sequences were found in 40 hypothetical proteins, suggesting they belong to secretory proteins and have membrane or extracellular localization.

During the analysis of protein fragments many oligopeptides were found, which had homologues in other microorganisms according to EROP_Moscow database. Considering the criteria of similarity 100-50% and physicochemical distance were selected 2567 oligopeptides in 1609 proteins. Different functional classes of homologues were noted and the ratio in proteins group "hypothetical" and "other" was comparable. The largest class – enzyme inhibitors (phosphatase in particular), which prevent the formation of phagolysosome. Another big group – pheromones (or their inhibitors) who control many important processes in prokaryotes, including cell-cell interaction. Moreover, a spectrum of probable proteases of substrain was detected.

Conclusion. The comparison of sequenced genomes vaccine substrains BCG-Russia and BCG-Tokyo demonstrated their high similarity. The bioinformatic analysis of proteome allows to predict hidden potential of hypothetical proteins, their involvement in metabolic pathways. The research of fragmentome showed that even after degradation of proteins their oligopeptides can participate in intercellular interaction in bacterial population and with host organism.

URL: http://rast.nmpdr.org – RAST. http://erop.inbi.ras.ru/ - EROP-Moscow.

ALIGNMENT OF «UNALIGNABLE» PROTEIN STRUCTURES

Aksianov E.A.*, Alexeevsky A.V.
Belozersky Institute of Physico-Chemical Biology, Moscow State University, Moscow, Russia
e-mail: evaksianov@gmail.com
**Corresponding author*

Key words: All-β- and α/β -proteins, protein topology, structural alignment

Motivation and Aim: Protein domains with the same fold according SCOP classification are supposed homologs. Nevertheless in many cases homology can be detected automatically neither by sequence nor structural alignments. At fold level, SCOP experts' decisions are mainly based on similarity of topology, i.e. spatial organization of secondary structure elements (SSE) along polypeptide chain. We developed a program detecting topological similarity between two protein structures of all- β and α/β-classes.

Definitions: Protein topology is encoded as described in [1]. For example, the code of PDB structure 1BWW (chain A) topology is a sequence of labeled strands

$$A_3^- - B_5^- - A_2^+ - B_2^- - B_1^+ - B_0^- - A_0^+ - A_1^- - B_3^+ - B_4^- \quad (1)$$

Strands are separated by "–", they are ordered from N- to C-terminus. Capital letters A and B denote sheets. Within each sheet strands are numbered form one (top) edge to other (bottom) sheet edge. Signs denote N to C strand directions within a sheet, either from "left" side to "right" side of the sheet or opposite. For example, "$A0^+ - A1^-$" is a β-hairpin. Sheet map (see http://mouse.belozersky.msu.ru/sheep) flips top-to-bottom and left-to-right result in changing topology code to equivalent ones. Renaming sheets and linear shift of numbering are also allowed. Equivalent sheet codes induce equivalent codes of whole structures. For example, (1) is equivalent to

$$A_1^- - B_0^- - A_2^+ - B_3^- - B_4^+ - B_5^- - A_4^+ - A_3^- - B_2^+ - B_1^- \quad (2)$$

Topology code (2) is obtained by flipping sheets A and B top-to-bottom and shifting numbering in sheet A by one.

Algorithm: Our program takes two PDB structures as input and find local alignment by dynamic programming algorithm. Heuristics are used to reduce exhausting of equivalent topology codes.

Results: Program detects topological similarity in accordance with human judgment. For example, the structure 1BWW is annotated as Immunoglobulin-like in SCOP. The structure 1MKF (chain A) belongs to another SCOP fold with announced similarity to immunoglobulins. The similarity between these structures cannot be established by sequence or structural alignment. Local topological alignment obtained by our tool is

$$A_3^- - B_5^- - A_2^+ - B_2^- - B_1^+ - B_0^- - A_0^+ - A_1^- - B_3^+ - B_4^-$$
$$D_1^- - h^2 - D_2^+ - C_3^- - C_4^+ - C_5^- - D_4^+ - D_3^- - C_2^+ - C_1^-$$

Topology similarity becomes evident if code (1) is replaced by equivalent code (2). Sheets A and D (B and C) corresponds to each other. Nine matches and one mismatch (strand `B5 -` vs helix `h2`) are detected. *Conclusion:* Developed alignment tool is useful for detecting weak similarity between highly diverged structural domains.

Availability: http://mouse.belozersky.msu.ru/protop.

Acknowledgements: this work was partially supported by RFBR grants 14-04-31709 and 13-07-00969.

References:

1. E. Aksianov. (2014) Motif Analyzer for protein 3D structures, J. Struct. Biol, **186:** 62-67.

BINDING SITES OF miRNA WITH MYB GENES' mRNA IN *B. TAURUS* AND *B. MUTUS*

Alybaeva A.Z.*, Niyazova R.E., Ivashchenko A.T.
National Nanotechnology Laboratory, *Al-Farabi KazNU, Almaty, Kazakhstan*
e-mail: aigul.alybaeva@mail.ru
**Corresponding author*

Key words: miRNA, mRNA, MYB family, binding site, animal

Motivation and Aim: MYB transcription factors (TF) family control the expression of genes involved in the cell cycle, apoptosis, metabolism, and other processes. Natural regulators of MYB genes expression are miRNAs, which may partially or completely inhibit the translation process. Action of miRNAs on animals' TF genes expression is not studied sufficiently, so it is important to ascertain the effect of miRNA on the expression of *MYB* TF family.

Methods and Algorithms: MirTarget defines the localization of miRNA binding sites in the 5'UTR, CDS and 3'UTR of the mRNAs, calculates the free energy hybridization (ΔG) and the ratio $\Delta G/\Delta G_m$ (%), where ΔG_m equal to the free energy of miRNA binding with total complementary nucleotide sequence. The sites were selected with ΔG/ΔGm more 85%.

Results: We studied the interaction of 749 miRNAs from *B.taurus* with mRNA of 20 *MYB* genes of *B.taurus* and 20 *MYB* genes of *B.mutus.* 58 miRNAs associate with eight *MYB* genes of *B.taurus* and 73 miRNAs associate with 18 *MYB* genes of *B.mutus.* It was revealed that five miRNAs act on mRNA of the *MIER1* gene of *B.taurus* and miR-1281, miR-2285 and miR-2337 bind with mRNA of the *MIER1* gene of *B.mutus*. mRNA of the *MYBL1* gene of *B.mutus* binds with seven miRNAs and three of them interact with the mRNA of the *MYBL1* gene of *B.taurus*. The binding sites for six miRNAs in mRNA of the *MYSM1* gene of *B.taurus* and *B.mutus* were established. mRNA of the *NCOR1* gene in *B.taurus* binds to 21 miRNAs and 18 of these miRNAs interact with mRNA of the *NCOR1* gene in *B.mutus.* 10 miRNAs act on the mRNA *RCOR1* gene of *B.taurus* and miR-1814b binds to *RCOR1* mRNA of *B.mutus.* mRNA of the *RCOR2* gene in *B.taurus* interact with seven miRNAs, five of them interact with mRNA of the *RCOR2* gene in *B.mutus.* With mRNA of the *SMARCC2* gene in *B.taurus* bind 14 miRNAs, and five of them bind with mRNA of *B.mutus.* mRNA of the *TTF1*gene in *B.taurus* binds with four miRNAs and only miR-2325 binds with mRNA gene in *B.mutus.* mRNA of the *MIER1* gene in *B.taurus* has eight binding sites for miRNA in CDS and one in 3'UTR. mRNA of the *MIER1* gene in *B.mutus* bind two miRNAs in CDS and one in 3'UTR. mRNA of the *MYB* gene in *B.mutus* has one binding site in CDS and other in 3'UTR. mRNA of the genes *MYBL1* and *MYSM1* in *B.taurus* and *B.mutus* bind with six miRNAs and their binding sites are located in 5'UTRs, CDSs and 3'UTRs. mRNA of the *NCOR1* in *B.taurus* has 21 binding sites for miRNAs in CDS and only one in 5'UTR. 10 miRNA binding sites with mRNA of the *RCOR1* gene in *B.taurus* are located: one in CDS, eight in 3'UTR and one in 5'UTR. Six binding sites mRNA of the *RCOR2* gene in *B.taurus* with seven miRNAs are located in CDS and one is located in the 3'UTR. mRNA of *the SMARCC2* gene in *B.taurus* bind with 14 miRNA, eight of them are located in 3'UTR and six are located in CDS. Seven miRNA have binding sites with mRNA of the *SMARCC1* gene in *B.mutus* in CDS and one in 3'UTR. mRNA of the *TTF1* gene in *B.taurus* has binding sites with four miRNAs, which are located in CDS and 5'UTR. Three miRNA binding sites with mRNA of the *TADA2A* gene in *B.mutus* are located in CDS and 3'UTR. 13 mRNA of the MYB genes in *B.mutus* and *B.taurus* have binding sites in the CDS.

Conclusion: The expression of some *MYB* genes TF family of *B.taurus* and *B.mutus* can be regulated by miRNA, which bind with mRNA *MYB* genes in CDS, 3'UTR and 5'UTR.

NATIVE PROTEINS AND DECOYS: SUBTLE STRUCTURE DIFFERENCES

Anashkina A.A.[1], Esipova N.G.[1], Kuznetsov E.N.[2], Tumanyan V.G.[1]
[1]Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow, Russia;
[2]V. A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences, Moscow, Russia
e-mail: anastasya.anashkina@gmail.com

Key words: protein structure, decoys, Voronoi-Delone tessellation

Motivation and Aim Decoys are an artificially created models of proteins. As a rule, modern decoys have native-like structures and are very close in energy to the native structures. However, such structures do not exist in nature. The ability to distinguish the native protein structure from decoys is very important for protein folding and protein structure prediction. Also it can be used for quality assessment of existing native structures.

Methods and Algorithms In this paper, we sought differences in the structures of the native proteins and decoys using Voronoi-Delaunay tessellation. Voronoi-Delaunay tessellation allows to determine unambiguously contacting atoms in space, and characterizes each contact by area size of appropriate edges of the Voronoi polyhedron [1] and by the distance between these atoms. The reciprocal of the volume of the Voronoi polyhedron characterizes the local density of atoms.

Results We have shown that differences in the histogram of contacts area are very small, but there are differences in the histograms of distances between the contacting atoms in native proteins and decoys. The main difference lies in the range 2.7-2.8 E. There are greatly less contacts in decoys than in native structures. Also we derived a statistical potentials based on contacts for decoys and native structures. Shape of the potential well of the native structure has a shallow wide bottom, while the shape of the potential well for decoys has narrow bottom like for van der Waals potential. Also there are differences in the histograms of contacts area between amino acids. The proportions of specific contacts for native structures and decoys were calculated.

Conclusion We have shown that a detailed analysis of the structures of native proteins and decoys by Voronoi-Delaunay tessellation method enables to differentiate between them.

Acknowledgements The research was partially supported by Presidium of Russian Academy of Science Programm “Molecular and Cell Biology”, grants 12-04-01776-A and 14-04-00639-A of Russian Foundation for Basic Research.

References

1. A.A. Anashkina, E.N. Kuznetsov, N.G. Esipova, V.G. Tumanyan. (2007) Comprehensive statistical analysis of residues interaction specificity at protein-protein interfaces, Proteins, 67(4):1060-77

COMPUTATIONAL PREDICTION OF NOVEL ANTI-HIV-1 AGENTS BASED ON POTENT AND BROAD NEUTRALIZING ANTIBODY VRC01

Andrianov A.M.[1], Kashyn I.A.[1], Tuzikov A.V.[2]*
[1]*Institute of Bioorganic Chemistry NAS Belarus, Minsk, Belarus;*
[2]*United Institute of Informatics Problems NAS Belarus, Minsk, Belarus*
e-mail: tuzikov@newman.bas-net.by
**Corresponding author*

Key words: HIV-1, gp120 protein, monoclonal antibody VRC01, virtual screening, molecular modeling, peptidomimetics, anti-HIV-1 therapeutics

Motivation and Aim: In light of discovering anti-HIV-1 broadly neutralizing antibodies (bNAbs) and mechanisms of their action, studies aimed at the identification of chemical compounds that are able to mimic pharmacophore properties of these antibodies are of great interest. In this work, computational prediction of novel HIV-1 entry inhibitors presenting peptidomimetics of potent and broad neutralizing antibody VRC01 was carried out followed by evaluation of their potential inhibitory activity by molecular modeling.

Methods and Algorithms: Computer-aided search for the most probable peptidomimetics of bNAb VRC01 was carried out based on the analysis of X-ray complex of this antibody Fab with the HIV envelope gp120 core. Using these empirical data, peptidomimetic candidates of bNAb VRC01 were identified by a public, web-oriented virtual screening platform (pepMMsMIMIC) and models of these candidates bound to gp120 were generated by molecular docking. Finally, the stability of the complexes of these molecules with gp120 was estimated by molecular dynamics and binding free energy simulations.

Results: The calculations identified six molecules exhibiting a high affinity to the HIV-1 gp120 protein. These molecules were selected as the most probable peptidomimetics of bNAb VRC01. In a mechanism similar to that of bNAb VRC01, these compounds were found to block the functionally conserved regions of gp120 critical for the HIV-1 binding to cellular receptor CD4. The docked structures of the identified molecules with gp120 do not undergo substantial rearrangements during the molecular dynamics simulations, exposing the low values of free energy of their formation.

Conclusion: Based on the above findings, the compounds selected are considered as promising basic structures for the rational design of novel, potent, and broad-spectrum anti-HIV-1 therapeutics.

Availability: The search for peptidomimetics was performed by the pepMMsMIMIC tools associated with the MMsINC database (free available at http://mms.dsfarm.unipd.it/pepmmsmimic.html). The AutoDock VINA program (free available at http://vina.scripps.edu/) was involved in the molecular docking calculations. The molecular dynamics simulations were carried out by the commercial package AMBER 11 (http://ambermd.org/).

PREDICTION OF ANTISENSE RNA-RNA INTERACTIONS IN ANIMAL CELLS

Antonov I.V.*, Marakhonov A.V., Baranova A., Skoblov M.Yu.
Research Centre for Medical Genetics RAMS, Moscow, Russia
e-mail: vanya.antonov@gmail.com
**Corresponding author*

Key words: antisense interactions, non-coding RNAs, RNA interactome, Staufen mediated decay

Motivation and Aim: Latest insights placed a novel emphasis on mutual interactions between two or more RNA molecules. Notably, a number of functional mRNA-mRNA and lncRNA-mRNA antisense pairs have recently been discovered [1-3]. Several thermodynamic-based tools were previously developed to predict potential targets for a query RNA [4, 5]. However, they are not suitable for transcriptome-wide identification of RNA targets – either due to time or accuracy restrictions. It is known that conserved regions of transcripts frequently bare specific biological functions. The goal of the current work is to develop an alternative approach to discover new functional antisense interactions that is based on sequence conservation and expression analysis.

Methods and Algorithms: The nucleotide version of BLAST is used for fast database screening.

Results: To decrease the total number of hits all the repeats (as identified by RepeatMasker, RM) and non-conserved regions (NCRs – the areas between annotated conserved elements) were masked in the target database. The masking approach was validated on a list of 31 well-studied sense-antisense human gene pairs (mRNA/mRNA or mRNA/lncRNA type). From each pair, one gene was used as a query. True Positive (TP) interaction was recorded, if the second gene of the same pair was predicted, while all non-target predictions were treated as False Positives (FP). When 31 validation queries were run against non-masked and RM+NCR masked RefSeq databases, the total numbers of retrieved hits (N_{hits}) were 55,741 and 1,677, respectively. Masking-related changes in sensitivity (Sn) and specificity (Sp) were 100% to 90.6% for Sn and 0.0057% to 1.75% for Sp. Posttranscriptional regulation is usually associated with 5' and/or 3' UTRs rather than CDS of the transcript. Additional CDS masking for protein coding genes resulted in further reduction of the number of hits and improvement of specificity (N_{hits}= 276, Sn = 50%, Sp = 5.8%, p-value = $1.5 \cdot 10^{-6}$). To find specific RNAs that regulate expression of other genes via antisense mechanism we analyzed publicly available single gene knockdown/overexpression microarray experiments. First, the antisense partners interacting via conserved regions were predicted. Next, for each experiment a subset of genes that changed their expression above threshold was selected and the fraction of such genes among the predicted partners was computed. Using hypergeometric distribution the p-value was calculated for each microarray experiment and the most promising RNAs were identified.

Conclusion: We have shown that a search using conserved parts of the transcripts decreases the number of possible antisense partners (hits) by several folds and significantly increases specificity. Analysis of microarray experiments revealed the most promising RNAs that may act via antisense mechanism.

Availability: the ASSA web server is available at: http://assa.generesearch.ru/

References:

1. Gong C, Tang Y, Maquat LE (2013) mRNA-mRNA duplexes that autoelicit Staufen1-mediated mRNA decay, Nat Struct Mol Biol, 10:1214-20.
2. Mahmoudi S, Henriksson S, Corcoran M, Mйndez-Vidal C, Wiman KG, Farnebo M. (2009) Wrap53, a natural p53 antisense transcript required for p53 induction upon DNA damage, Mol Cell. 33(4): 462-71.
3. Yoon JH el al. (2012) LincRNA-p21 suppresses target mRNA translation. Mol Cell, 47(4):648-55
4. Tafer H, Amman F, Eggenhofer F, Stadler PF, Hofacker IL (2011) Fast accessibility-based prediction of RNA-RNA interactions. Bioinformatics, 27(14):1934-40.
5. Wenzel A, Akbasli E, Gorodkin J. (2012) RIsearch: fast RNA-RNA interaction search using a simplified nearest-neighbor energy model. Bioinformatics, 28(21):2738-46.

INTRON LENGTHS AND PHASES: REGULARITIES AND DATABASE

Astakhova T.V. [1] , Tsitovich I.I. [2, 3], Yakovlev V.V. [1, 3], Roytberg M.A. [1, 3, 4]*
[1]*Institute of Mathematical Problems of Biology, RAS, Moscow Region, Russia;*
[2]*Institute of Information Transmission Problems, RAS, Moscow, Russia;*
[3]*Higher School of Economics, Moscow, Russia;*
[4]*Moscow Institute of Physics and Technology, Moscow Region, Russia*
e-mail: mroytberg@lpm.org.ru
** Corresponding author*

Key words: exon, intron, intron phase, intron length

Motivation and Aim: Understanding of regularities of exon-intron structure is important for study of mechanisms and evolution of exon-intron structure. Relationships between phase and length of introns are not sufficiently studied. We extend our previous results [1] to genomes of Plantae.

Methods and Algorithms: Phase of intron is a remainder of total length of preceding exons divided by three. In [1] we have considered 17 animal genomes of various classes (insects, fishes, amphibians, reptiles, birds, mammals) and demonstrated for all the genomes the following effect. It is known that introns are unevenly distributed over the phases, the phase ratio, i.e. proportion between number of introns in the phases 0, 1 and 2, is about 5:3:2. We have shown that the ratio changes, if we consider only T-long introns, i.e. introns, with length greater than a threshold T. With increasing of the threshold T part P1(T) of introns in phase 1 increases, while the part P0(T) of introns in phase 0 decreases and at a certain value we obtain P1(T)=P0(T). In this work we extend the study to 4 plant genomes, *A.thaliana, O.sativa, V.vinifera* and *P. trichocarpa*.

Results: For all considered plant genomes the baseline ratio is about 6:2:2, i.e. fraction of phase 0 introns in plant genomes is greater than it in animal genomes. For *V. vinifera* and *P. trichocarpa* the ratio P0(T):P1(T):P2(T) unlike all other considered genomes does not depend on T. However for *O. sativa* and *A. thaliana* the dependence exists. E.g. for *O. sativa* with T=1100 the phase ratio is 53:23:24 while for all introns we have ratio 58:20:22. Z-score ZP0(1100) of the decrement of P0(1100) compared to P0(0) is -5.4. Number of introns having length at least 1100 is 2915 that is 5.6 % of all introns. For *A. thaliana ZP0(T)* has two local extrema. For T=1200 we have ZP0(1200) = -5.6, number of long introns 132 (0.2 % of all introns) and phase ratio 34:27:29. For T=280 we have ZP0(280) = -4.0, number of long introns 6015 (9.2% of all introns) and phase ratio 55:21:23. Some other regularities related to intron lengths and phases were also revealed. To provide the study we have created the database available at http://server2.lpm.org.ru/static/introns_db/ ; the scheme of the database can be found at the same site.

Conclusion. The dependence between lengths and phases in plant genomes significantly differs from those for animals. That may reflect difference in evolution of the taxa.

References:

1. T.V. Astakhova, I.I. Tsitovich, V.V. Yacovlev, M.A. Roytberg. Dependence between lengths and phases of introns.// *MCCMB 2013: Proceedings*. 2013.

APPLICATION OF TERMINAL HELITRON FRAGMENTS AS HIGH POLYMORPHIC MARKERS OF GENOME SCANNING IN UNGULATA

Barducov N.V.*, Sipko T.P., Glazko V.I.
Russian State Agrarian University – Moscow Timiryazev Agricultural Academy
e-mail: nikolay_bardukov@yahoo.com
**Corresponding author*

Key words: Helitrons, multiloci genome scanning, *Ovibos moschatus*, *Equus caballus*

Motivation and aim: A search of highly polymorphic markers for multiloci genome scanning is a very important task for solution of many problems such as a control of gene pool dynamics of populations, an estimation of breeds consolidation and an estimation of phylogenetic relationships. The application of Helitrons' terminal flanks as highly informative markers for genomic scanning was investigated in the present study. Helitrons are a special type of DNA transposons which replication method is named "rolling-circle" mechanism.

Methods and algorithms: The primers were selected to terminal flanks of Helitrons from the *Heligloria* family which is widely distributed in mammalian genomes.

Results: The polyloci genome scanning of ancient and modern reintroduced populations of muskoxen and also Altaic and Karachay horse breeds was performed. Phylogenetic relationships between the investigated groups were identified. The assessment of animal groups consolidation was carried out using selected markers. A possible relationship between palindrome region within mobile genetic elements of current class and their polymorphism is discussed.

Conclusion: The data acquired evidence that mammal genomes may include nucleotide sequences with homology to Helitrons. These sequences may be localized in short distances in different DNA strands and their application as PCR primers produce polymorphic multiloci amplification spectra which can be used in gene pool dynamics control. We observed coincidences in amplification spectra of modern and ancient muskoxen as well as differences of such spectra in investigated horse breeds. This fact allows supposing some relation between animal adaptation processes and distribution features of regions with homology to Helitrons.

Availability: High polymorphic molecular markers of multiloci genome scanning have good chances in application as for gene pool dynamics control among wild species as in agricultural selection and breeding.

Acknowledgements: We would like to acknowledge the museum "Our Ice Age" for samples of ancient muskoxen. This work is done with support of RFBR grant № 13-04-02199 A.

ANALYSIS AND CLASSIFICATION OF NONSTANDARD RNA MOTIFS

Baulin E.F.[1, 2]*, Spirin S.A.[3] , Roytberg M.A.[1, 2, 4]
[1]*Institute of Mathematical Problems in Biology, Pushchino, Russia;*
[2]*Higher School of Economics, Moscow, Russia;*
[3]*Belozersky Institute of Moscow State University, Moscow, Russia;*
[4]*Moscow Institute of Physics and Technology, Moscow Region, Russia*
e-mail: baulin@lpm.org.ru
**Corresponding author*

Key words: RNA structure, database, nonstandard motif, pseudoknot, multiplet

Motivation and Aim: Investigation of RNA and its spatial structure is one of the main areas of modern molecular biology, both in theoretical (understanding of intracellular processes, the theory of evolution), and in practical parts (drug discovery). Development of research is constrained by the fact that there is no universal annotated database of spatial structures of RNA. One of the reasons for that is the lack of a common classification of RNA structure elements, which would allow one to describe both classical secondary structures and non-standard structures. In our previous work [1] we have proposed the universal classification of RNA structure elements; the database of spatial structures of RNA based on this classification was developed [2]. In this work we study structures formed by interactions between RNA helices (*multiplets*), and some other non-standard RNA structure motifs, e.g. pseudoknots.

Methods and Algorithms: RNA-containing structures were selected from Protein Data Bank [3] and then base-pairs were annotated using new program DSSR [4]. These two types of data were used as input of original program package to mine new data and create the database [2]. The database architecture is based on the original classification of RNA secondary structure elements.

Results: Classification of multiplets is based on their graph representation; the vertices correspond to helices, the edges reflect H-bonds between helices. Most common multiplets are formed by two helices (12772 cases). The complete list of multiplet types that occur in PDB RNA structures augmented with all occurrences of multiplets was created. The list of all types of multiplets includes 75 types, the most complicated multiplet contain 12 helices (see structure with PDB code 3UZM); a helix may be connected with up to 5 other helices (see structures 3UZM, 1FCW and 3KC4).

Conclusion: We have demonstrated usefulness of the created database and proposed classification of non-standard RNA structures.

Availability: The beta-version of the database is available at http://server2.lpm.org.ru/~baulin/home.html.

References:

1. Баулин Е. Ф., Астахова Т. В., Ройтберг М. А. //*Математическая биология и биоинформатика*. – 2012. – Т. 7. – №. 2. – С. 567-571.
2. E.F. Baulin, M.A. Roytberg. // *MCCMB 2013: Proceedings*. 2013.
3. Berman H. M. et al. //*Nucleic acids research*. – 2000. – Т. 28. – №. 1. – С. 235-242.
4. Lu X. J., Olson W. K., Bussemaker H. J. //*Nucleic acids research*. – 2010. – Т. 38. – №. 14. – С. 4868-4876.

A MACHINE LEARNING ANALYSIS OF URINE PROTEOMICS IN SPACE-FLIGHT SIMULATIONS

Binder H.[1], Wirth H.[1], Arakelyan A.[2], Lembcke K.[1], Tiys E.S.[3], Ivanishenko V.[3], Kolchanov N.A.[3], Kononikhin A.[4], Popov I.[5], Nikolaev E.N.[4, 5], Pastushkova L.[6], Larina I.M.[6]

[1]*Interdisciplinary Centre for Bioinformatics, Universität Leipzig, Leipzig, Germany;*
[2]*Institute of Molecular Biology NAS RA; Yerevan, Armenia;*
[3] *Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia;*
[4] *Institute for Energy Problems of Chemical Physics, RAS, Moscow, Russia;*
[5]*Emanuel Institute for Biochemical Physics, RAS, Moscow, Russia;*
[6]*Institute of Biomedical Problems – Russian Federation State Scientific Research Center RAS, Moscow, Russia*

Key words: Mars 105 space-flight experiment, urine protein expression, salt balance, molecular processes, bioinformatics analysis, self organizing maps, visualizing high throughput data

Long-term space travel simulation experiments enabled to discover new aspects of human metabolism. Detailed proteomics data were collected during the Mars105 isolation experiment potentially enabling a deeper insight into the molecular processes involved. Machine learning using self organizing maps (SOM) in combination with different analysis tools was applied to describe the time trajectory of protein expression. The method portrays the protein expression landscapes enabling a personalized and intuitive view on the physiological state of the volunteers. The dynamics of urine proteomics was described in terms of trajectories reflecting the time evolution of the protein expression landscapes with individual resolution. The abundance of more than one half of the proteins measured clearly changes in the course of the experiment. The trajectory splits roughly into three time ranges, an early (week 1-6), an intermediate (week 7-11) and a late one (week 12-15).The total protein expression level is maximum in the early time region and then it progressively decreases until the end of the experiment. Regulatory modes associated with distinct biological processes were identified using previous knowledge by applying enrichment and pathway flow analysis. Early protein activation modes can be related to immune response and inflammatory processes, activation at intermediate times to developmental and proliferative processes and late activations to stress and responses to chemicals. These protein expression profiles support previous results about alternative mechanisms of salt storage in an osmotically inactive form paralleled by the activation of immune responses in the context of micro-vascularization. Our study shows that SOM machine learning in combination with analysis methods of class discovery and functional annotation enable the straightforward analysis of complex proteomics data sets generated by means of mass spectrometry.

FLUORESCENCE *IN SITU* HYBRIDIZATION WITH MICRODISSECTED DNA PROBES ON CHROMOSOMES OF SPECIES WITH LARGE GENOME SIZE WITHOUT SUPPRESSION OF REPETITIVE DNA SEQUENCES

Bogomolov A.G.[1]*, Jetybayev I.E.[1], Podkolodnyy N.L.[1, 2], Rubtsov N.B.[1, 3]
[1]*Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia;*
[2]*Institute of Computational Mathematics and Mathematical Geophysics SB RAS, Novosibirsk, Russia;*
[3]*Novosibirsk State University, Novosibirsk, Russia*
e-mail: mantis_anton@bionet.nsc.ru
**Corresponding author*

Key words: fluorescence in situ hybridization (FISH), chromosomal in situ suppression hybridization (CISS- hybridization), image analysis

Motivation and Aim: Suppression of repetitive DNA sequences [1] is a standard use in chromosome painting with the DNA probes derived from an individual chromosome. However, for species with large genomes CISS-hybridization doesn't significant improve hybridization signals. The main reason for this is considered to be an inefficient blocking of repetitive sequences. One of possible approaches to overcome the described problem would be the usage of the computer method [2]. We refer to this approach as VISSIS (**Vi**sualization **s**pecific **s**ignals ***i****n* ***s****ilico*). Here we present results of VISSIS applied to grasshopper chromosomes.

Methods and Algorithms: Grasshoppers were used as species with relatively large genomes. In this study DNA probes were generated by microdissection of sex chromosomes of *Nocaracris cyanipes*, *Paranocaracris rubripes* and *Paranothrotes opacus* from Pamphagidae family following DOP-PCR. To detect specific FISH signals, which allow to identify chromosomes and theirs regions, an improved VISSIS method was used. In order to analyze the signal distribution in autosome the following additional procedures were applied: separate processing of bright regions in the image (excluding of C-positive blocks), classification of image signals according to the level of intensity (in order to identify the signals from DNA sequences specific only for chromosomes or their regions, which DNA probes were derived from).

Results: The method was tested on images of CISS-hybridization and FISH with metaphase and meiotic chromosomes.

Conclusions: The results obtained showed, that the method in general allowed identifying target chromosome and chromosome regions without suppression of repetitive DNA sequences. More over it could be used for improving results of CISS-hybridization (especially in case of incomplete suppression). Notice that 'chromosome painting' in species with relatively large genomes could not be carried out before.

Acknowledgements: This study was supported by grant from OPTEC LLC.

References:

1. Lichter P. et al. (1988) Delineation of individual human chromosomes in metaphase and interphase cells by in situ suppression hybridization using recombinant DNA libraries, Human genetics, 80: 224-234
2. Bogomolov A. G. et al. (2012) Visualization of chromosome-specific DNA sequences by fluorescence in situ hybridization of microdissection DNA probes with metaphase chromosomes, Russian Journal of Genetics, 2: 413-420

WHOLE-GENOME COMPARATIVE ANALYSIS OF CpG ISLANDS BETWEEN HUMAN AND CHIMPANZEE

Borzov E.A.[1], Mehta R.[3], Baranova A.V.[1, 3], Skoblov M.Yu.[1, 2 ,4]*
[1]*Research Centre for Medical Genetics RAMS, Moscow, Russia;*
[2]*State Budgetary Institution of Higher Education Moscow State Medical and Dental University, Moscow, Russia;*
[3]*School of Systems Biology, College of Science, George Mason University, Fairfax, VA USA;*
[4]*The Russian National Research Medical University named after N.I. Pirogov (RNRMU)*
e-mail: mskoblov@generesearch.ru
**Corresponding author*

Key words: CpG island, evolution, conservation

Motivation and Aim: The study of genomic sequences from different organisms has convincingly shown that one of the main factors of evolution is a change in the regulation of genes, rather than mutations in their protein-coding sequences as previously thought. Among the large variety of regulatory mechanisms, epigenetic mechanisms are the main ones. Methylation of CpG islands is a powerful mechanism for fine regulation involved in a lot of processes in normal and in pathological conditions. The aim of this work was to investigate the conservatism of CpG islands in human and chimpanzee.

Results: We have performed the whole-genome comparison of CpG islands in human and chimpanzee and revealed three groups of CpG islands: diverged, medium and extremely conserved. We have characterized these groups and identified genes which can regulate these islands. Identified genes were functionally characterized using Metacore software. We have also conducted a comparative genomic analysis of CpG islands in human and HomoDenisova.

Conclusion: We have shown that about 10% CpG islands belong to a group of highly conserved sequences, despite the fact that the genome had a very strong selection against CpG dinucleotides. We have also found that 4% of CpG islands have a very wide variety in sequences.

References:

1. Fukuda K, Ichiyanagi K, Yamada Y, Go Y, Udono T, Wada S, Maeda T, Soejima H, Saitou N, Ito T, Sasaki H. Regional DNA methylation differences between humans and chimpanzees are associated with genetic changes, transcriptional divergence and disease genes. J Hum Genet. 2013 Jul;58(7):446-54.
2. Schneider E, El Hajj N, Richter S, Roche-Santiago J, Nanda I, Schempp W, Riederer P, Navarro B, Bontrop RE, Kondova I, Scholz CJ, Haaf T. Widespread differences in cortex DNA methylation of the “language gene” CNTNAP2 between humans and chimpanzees. Epigenetics. 2014 Jan 16;9(4).

COMPUTER ANALYSIS OF EXPRESSION LEVEL OF ALLERGEN-CODING GENES OF PATHOGENIC MICROORGANISMS

Bragin A.O.*, Sokolov V.S., Demenkov P.S., Matushkin Yu.G., Ivanisenko V.A.
Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
e-mail: ibragim@bionet.nsc.ru
**Corresponding author*

Key words: protein allergenicity prediction, allergenicity of microbial proteomes, habitat of microorganisms

Allergy is one of the most widespread health concerns in our days. More than one third of the world population suffers from allergic diseases. Despite progress in the study of the etiology of allergic diseases the effect of various microorganisms on IgE-mediated reactions in human remains poorly studied. Knowledge about the relationship of proteins of microorganisms with IgE-mediated allergy is important for medicine since elimination therapy is one of the most effective ways to prevent an appearance of symptoms of allergy. In this regard arises the need for analyze allergenic properties of bacteria proteins.

Genes and their encoded proteins of 587 species of bacteria and archaea were taken from the BioProject and BioSample databases. Microorganisms that cause disease in mammals we considered as pathogenic. Gene expression level was estimated by the method proposed by Likhoshvai *et al* 2002. Predicting allergenicity of proteins was carried out by the method developed Bragin *et al* 2012. The evaluation of relationship between pathogenicity of microorganisms, the level of gene expression and allergenicity of protein was performed using the Chi-square method.

By using computer methods in this work was carried out analysis the correlation of gene expression with the level of allergenicity of the encoded proteins. It has been found that the level of gene expression was significantly higher in the pathogenic microorganisms. Furthermore, it was shown an increased number of highly expressed genes among the genes encoding allergenic proteins.

References:

1. Likhoshvai V. A., Matushkin Y. G. Differentiation of single-cell organisms according to elongation stages crucial for gene expression efficacy //FEBS letters. – 2002. – T. 516. – №. 1. – C. 87-92.
2. Bragin A. O., Demenkov P. S., Ivanisenko V. A. Protein allergenicity prediction on the basis of conformational peptides //Russian Journal of Genetics: Applied Research. – 2012. – T. 2. – №. 1. – C. 18-22.

PACKAGE OF FUNCTIONS FOR SCRIPT PROGRAMMING LANGUAGE R FOR TESTING THE CONVERGENCE OF POPULATION SAMPLES OF DNA SEQUENCES

Bukin Yu.S.
Limnological Institute SB RAS, Irkutsk, Russia
e-mail: bukinyura@mail.ru

Sample DNA sequences are widely used for population genetic analysis. With the help of DNA sequences set population structure of species, calculated various population parameters and gene flow between isolated groups of organisms. Another important aspect of the use of population samples of DNA sequences is to establish the demographic history of populations and species. In the recently developed DNA barcoding for species identification. The DNA barcoding requires precise knowledge of the interspecies and intraspecies polymorphism. All of these techniques as well as any method of statistical analysis requires a representative sample of processed data. In our case, is required to estimate the representativeness of the sample DNA sequences.

Most often in the assessment of a representative sample of DNA sequences evaluate the convergence of the calculated parameter - the average genetic distance between s sequences of samples. Assessment of convergence average genetic distance produced using bootstrap analysis. These techniques are incorporated in some programs, such as Arlequin311 [2] and DNASP [3]. With the help of this programs, you can calculate the percentage error which is determine on average genetic distance to specify a set of DNA. All of these programs have a number of drawbacks. First of inconvenience is to use graphical user interface, which makes use of programs in complex scripts for data processing. The second problem is due to the fact that these programs do not check for the convergence of the average genetic distance estimated for the two compared sets of DNA. It does not give the opportunity to use these programs in DNA barcoding process after comparing intraspecies and interspecies polymorphism.

In our work, we propose a generic package function for R programming language that with the aid of bootstrap methods define convergence sets of DNA sequences. The package contains functions for evaluating the convergence of genetic distances within and between samples. Reading sequences in FASTA format occurs with the functions of package APE [1] for R. Offers functions can include complex scripts for data analysis in the language R.

Testing and demonstration of a set of function carried out with the help of artificially simulated data samples of DNA in the program GENPOP [4]

References:

1. E. Paradis, J. Claude, K. Strimmer (2004). APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20: 289-290.
2. Excoffier, L.G. Laval, S. Schneider (2005) Arlequin ver. 3.0: An integrated software package for population genetics data analysis. Evolutionary Bioinformatics Online 1:47-50.
3. P. Librado, J.Rozas, (2009). DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25:** 1451-1452.
4. A. Carvajal-Rodríguez GENOMEPOP: A program to simulate genomes in populations. BMC Bioinformatics. 2008. V. 9

DRIVING FORCE FOR PROTEIN FOLDING: THE TWO-COMPONENT POTENTIAL

Chekmarev S.F.
Institute of Themophysics SB RAS, Novosibirsk, Russia;
Novosibirsk State University, Novosibirsk, Russia
e-mail: chekmarev@itp.nsc.ru

Key words: free energy surface, folding flow, source and sink of the flow, canalization of the flow

Motivation and Aim: To gain insight into the protein folding reaction, the free energy surfaces (FESs) are usually considered. However, the FESs define only the probability for the system to visit protein states. They do not show the direction in which the system proceeds, i.e., the protein can have the same probability to be in some state when it goes towards the native state or requires partial unfolding to reach the native state.

Methods and Algorithms: Following the hydrodynamic approach to protein folding [1], the fluxes of transitions between the protein states in a two-dimensional space of collective variables are calculated on the basis of simulated folding trajectories of the protein. In this approach, the folding process can be viewed as a steady flow of the representative points of the protein from the unfolded to native state. The transition to the two-dimensional space is performed by reducing the multi-dimensional conformation space with the Principal Component Analysis method. Using the Helmholtz decomposition of the vector field of folding fluxes in the reduced space of collective variables, a potential of the driving force for protein folding is calculated [2].

Results: It has been shown that the potential of driving force has two components. One component is responsible for the source and sink of the folding flows, which represent respectively, the unfolded states and the native state of the protein, and the other, which accounts for the flow vorticity inherently generated at the periphery of the flow field, is responsible for the canalization of the flow between the source and sink. The theoretical consideration is illustrated by calculations for a beta-hairpin protein.

Conclusion: The one-component FESs that are often referred to as the "potentials of mean force" (e.g., Ref. [3]) do not actually represent the potentials of driving force for protein folding. The corresponding ("true") potentials have two-components: one component accounts for the source and sink of the folding flow, and the other for the canalization of the flow between the source and sink.

References:

1. S.F. Chekmarev, A.Yu. Palyanov, M. Karplus. (2008) Hydrodynamic Description of Protein Folding, *Phys. Rev. Lett.,* **100:** 018107 (1-4).
2. S.F. Chekmarev. (2013) Protein folding: Complex potential for the driving force in a two-dimensional space of collective variables, *J. Chem. Phys.*, **139**: 145103 (1-8).
3. J.-E. Shea, C.L. Brooks III. (2001) From Folding Theories to Folding Proteins: A Review and Assessment of Simulation Studies of Protein Folding and Unfolding, *Annu. Rev. Phys. Chem.* **52**: 499-535.

BIOPHYSICAL PRINCIPLES GUIDING NUCLEOSOME POSITIONING *IN VIVO*

Chereji R.V.[1, 2], Kan T.-W.[3], Guryev V.P.[4], Morozov A.V.[1, 2], Moshkin Y.M.[3, 5]*
[1]*Department of Physics and Astronomy, Rutgers University, Piscataway, New Jersey, USA;*
[2]*BioMaPS Institute for Quantitative Biology, Rutgers University, Piscataway, New Jersey, USA;*
[3]*Erasmus Medical Center, Rotterdam, the Netherlands;*
[4]*European Research Institute for the Biology of Ageing, University of Groningen, Groningen, the Netherlands;*
[5]*Center of Genetic Resources, Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia*
e-mail: i.mochkine@erasmusmc.nl
**Corresponding author*

Motivation and Aim: Cells continuously adjust their transcriptional outputs in response to cellular and metabolic cues. Eukaryotic transcription, however, is interfaced with chromatin. Nucleosome is a basic building block for chromatin comprising 147 bp of DNA tightly wrapped around a core histone octamer. A major consequence of the packaging of genomic DNA into chromatin is that nucleosomes can impede access of DNA-binding proteins to their sites. To counter this, eukaryotic cells evolved complex ATP-dependent chromatin remodelling machineries (remodelers).

Results: The prevailing view in the field is that ATP-dependent chromatin remodelers affect nucleosome structure and positioning by disrupting contacts between the histone octamer and DNA, which results in nucleosome sliding, eviction, or histone replacement. Remodelers are highly abundant in cells, and our previous work has shown that they are capable of affecting nucleosome positions globally [1]. Surprisingly, we found that in *Drosophila* S2 cells there are no enrichment of remodelers at promoters of either active or silenced genes. Nonetheless, both inter-nucleosome spacing and positioning downstream of transcription start sites (TSS) are affected. To resolve this paradox, we turned into biophysical models, which show that histone concentration and turnover rates directly affect inter-nucleosome spacing [2]. Thus, we hypothesize that in addition to reposinioning nucleosomes via specific interactions, remodelers modulate nucleosome spacing over genic regions through direct or indirect regulation of histone gene expression levels, which is substantiated by our results.

Conclusion: These results lead to a paradigm-changing idea: that in addition to directly targeting a fraction of genomic nucleosomes, chromatin remodelers change nucleosome positions and occupancies globally simply by modulating availability of histone proteins and their turnover rates.

Acknowledgements: **YMM** is profoundly grateful to **Andrew Travers** for stimulating ideas, **Mikhail P. Moshkin** for his help with interpretation of the results from a broad biological perspectives and **C. Peter Verrijzer** for his introduction into the subject and financial support to the project.

References:

1. Moshkin Y.M., *et al.* (2012) Remodelers Organize Cellular Chromatin by Counteracting Intrinsic Histone-DNA Sequence Preferences in a Class-Specific Manner. *Mol Cell Biol* 32(3):675-688.
2. Chereji R.V. & Morozov A.V. (2011) Statistical Mechanics of Nucleosomes Constrained by Higher-Order Chromatin Structure. *J Stat Phys* 144(2):379-404.

HIGH-PERFORMANCE COMPUTING PROVIDES INSIGHT INTO THE INNERMOST ORGANIZATION OF PROCARYOTIC MEMBRANES

Chugunov A.O.[1]*, Volynsky P.E.[1], Efremov R.G.[1]
[1]*Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry RAS, Moscow, Russia*
e-mail: batch2k@yandex.ru
**Corresponding author*

Motivation and aim: Each domain of life possesses individual lipid composition of plasma membranes. *Bacteria* and *Eucaryota* feature zwitterionic and anionic phospholipids (PLs), respectively, organized in the bilayer manner. *Archaea* exhibit "bolalipid" monolayer membrane with each molecule possessing two O- or Θ-linked isoprenoid chains and two polar "heads" that form opposite hydrophilic membrane surfaces. Most probably, individual structural plan of the membrane determines unique ecological portrait of *Archaea*, which are extremophiles, and plays an important role in growth and division of *Bacteria*. Procaryotic membranes appear to be a unique target for antibacterials, in case of pathogenic bacterial strains, and promising bionanotechnological material, in case of *Archaea*. Here, we study the fine dynamic organization of bacterial and archaeal membranes with their microenvironment using high-performance molecular dynamics (MD) simulations.

Methods and Algorithms: The typical size of MD cell is ≈15×15×15 nm^3, which contains ≈5–20×10^4 atoms of lipids (, protein) and solvent. Accessible simulation time for common supercomputers is 0.5–1 μs, which is already long enough for model biomembranes to express their dynamic features. Our "bacterial" membrane was anionic and consisted of 75 % POPG and 25 % POPE PLs, with additional lipid-II molecule, which is the key player in bacterial cell wall synthesis and stands for promising target for novel antibacterial therapy. "Archaeal" membrane was built of a series of bolalipid "mimetics" with two hydrophobic chains, connected to polar "heads" at each end. In this series, acyl chains contained different number of characteristic for *Archaea* methyl or cyclopentanyl groups. Analysis of MD trajectories revealed local heterogeneities in bacterial membrane with lipid-II and complex phase behavior of archaeal membranes.

Results: For bacterial membrane, our analysis revealed that lipid-II molecule substantially disturbs lipid bilayer (as compared to the membrane without lipid-II) and introduces unique amphiphilic pattern into the membrane surface. Potentially, this complex pattern constitutes the "moving target" for many natural antibiotics, and may be employed in design and/or discovery of novel antibiotics against drug-resistant pathogenic strains.

For archaeal membranes, MD uncovered the role of "side" groups in bolalipids' hydrophobic chains: "straight" acyl tails form completely rigid "gel" phase, which is incompatible with life. Introduction of "side" groups gradually "melts" the membrane, finally arriving to "liquid crystal" state, which is believed to be native for the membrane. Such membrane composition may have adjusted *Archaea*'s ecological favor and evolutionary fate.

Conclusion: Molecular modeling is a powerful experimental toolbox for study of innermost details of biological systems, and its capabilities and role in near future will continue to increase. It fruitfully supplements traditional biophysical methods of research.

Acknowledgements: Supported by 13-04-40326-H and 14-04-3163414 RFBR grants.

MOLECULAR DYNAMICS STRUCTURE MODELING OF IL-36 CYTOKINES

Davidovich P.[1]*, Aksenova V.[1], Martin S.J.[1, 2]
[1]*St. Petersburg State Technological Institute, St. Petersburg, Russia;*
[2]*Trinity College, Dublin 2, Ireland*
e-mail: davidovich.pavel@technolog.edu.ru
**Corresponding author*

Key words: interleukin-36 proteins, structure modeling, molecular dynamics

Motivation and Aim: IL-36 family cytokines have been implicated to play a key role in skin inflammatory diseases, such as psoriasis. Upon activation, IL-36 cytokines (alpha, beta and gamma) interact with the IL-36 receptor (IL-36R) and initiate the production of a range of pro-inflammatory cytokines and chemokines. Similar to most members of the IL-1 family [1], IL-36 cytokines require proteolytic processing for activation. Thus, one therapeutic approach for blocking IL-36 activation is to inhibit the proteases responsible for activating IL-36 family cytokines. An alternative approach is to antagonize the interaction between IL-36 and the IL-36R. However, little is known concerning how IL-36 interacts with its cognate receptor. 3D-models of IL-36 cytokines are required to investigate the PPI between processed interleukins, proteases and IL-36R, but no structural data is so far resolved for IL-36 proteins.

In the current study we have studied the behavior and intramolecular interactions of monomeric IL-36 proteins. For that purpose we have built models for full-length IL-36 alpha (158 aa), IL-36 beta (164 aa) and IL-36 gamma (169 aa). We have accrued out a 500 ns-long molecular dynamics (MD) simulation to validate the stability of the FL models and predict their structural and dynamic properties.

Methods and Algorithms: The models of the full-length IL-36 cytokines were built using the I-TASSER server service from the human amino acid sequence. IL-1 beta template structure was generally used for threading. The protein structure predictions were made with medium confidence for the realistic starting structure to be sampled by the MD simulation. The starting structure has been immersed in a periodic box of TIP3P water model. The PME method was used to treat the long-range electrostatics. A time step of 2 fs was used. The conformational sampling was done at a temperature of 300 K using the V-Rescale algorithm. The equilibration procedure involved two rounds of minimizations and dynamics with decreasing force constants on the solute and protein atoms an unrestrained minimization of the whole system.

Results: We have built structural models of three IL-36 proteins (alpha, beta and gamma) and their processed forms by a combination of protein threading and MD simulation.

Conclusion: Structural information for IL-36 proteins is essential for the investigation of PPI with proteases responsible for IL-36 processing and IL-36 receptor interactions. Understanding these interactions is essential for drug design in the field of inflammatory diseases.

Availability: none

Acknowledgements: The work has been supported by Ministry of Education and Science of the Russian Federation State Contract № 14.B25.310013.

References:

1. I.S. Afonina et al. (2011) Granzyme B-Dependent Proteolysis Acts as a Switch to Enhance the Proinflammatory Activity of IL-1α, *Molecular Cell,* **44**: 265-278.

SPECTRAL CHARACTERISTICS OF TRYPTOPHAN IN WATER AND *VIBRIO HARVEYI* LUCIFERASE

Deeva A.A.*, Nemtseva E.V., Kratasyuk V.A.
Siberian Federal University, Krasnoyarsk, Russia;
Institute of Biophysics SB RAS, Krasnoyarsk, Russia
e-mail: anya.a.koval@gmail.com
**Corresponding author*

Key words: tryptophan fluorescence, protein intrinsic fluorescence, DFT, TD-DFT

Motivation and Aim: The aim of the study is to obtain the absorption and fluorescence spectra of tryptophan (Trp) aminoacid in aqueous solution and individual Trp residues in *Vibrio harveyi* (*V.h.*) luciferase using theoretical methods. The fluorescence analysis of experimental data on dynamics and conformational transitions of this protein in different conditions is complicated due to its multi-tryptophan fluorescence. The combination of theoretical and experimental study will shed light on the inner dynamics of *V.h.* luciferase.

Methods and Algorithms: The crystal structure of *V.h.* luciferase (PDB entry: 3FGC) and Trp (L-tryptophan ChemSpider ID: 6066) were studied.The DFT and TD-DFT methods were used to investigate the ground-state and excited-state properties, respectively. In the present calculations B3LYP/6-31+G(d,p) and B3LYP/3-21G level of theory was chosen for the optimization calculation of Trp in solution and in the protein, respectively. No constraints were used to the Trp atoms and the closest to indole ring polar atoms and hydrogens, the other atoms were kept fixed during the optimization. PCM was used to model a water solution and PCM with a dielectric constant of 4.0 was used to represent the protein environment. The calculations were carried out using GAMESS program package [1].

Results: To explain the experimental absorption and fluorescence spectra of Trp in water a zwitterionic form was studied. Trp residues from *V.h.* crystal structure were cut along with the residues located at the distance of ~5.0E from Trp atoms for the optimization procedure. The absorption energies and fluorescence energy of Trp zwitterion and Trp residues surrounded by the protein environment were obtained. The calculated spectral features of Trp in water are in good agreement with the experimental results. The obtained fluorescence peaks for Trp residues from *V. h.* allowed to assign them to the five spectral-structural classes [2].

Conclusion: In the present study the spectral characteristics of Trp in aqueous solution and Trp residues from *V. h.* were investigated. An approach of the structural changes, substrates or quenchers binding mechanisms analysis is proposed.

Acknowledgements: The research was partly supported by the Project 1762 from The Ministry of Education and Science of the Russian Federation.

References:

1. M.W. Schmid et al. (1993) *J. Comput. Chem.*, **14:** 1347-1363.
2. E.A. Burstein. (1983) The intrinsic luminescence of proteins is a method for studies of the fast structural dynamics, *Molecular Biology,* **17:** 455-467.

ASSEMBLING GENOMES AND METAGENOMES USING CLUSTER ARCHITECTURE

Demenkov P.S.[1, 2]*, Ivanisenko V.A.[1]
[1]*Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia;*
[2]*Novosibirsk State University, Novosibirsk, Russia*
e-mail: demps@bionet.nsc.ru
**Corresponding author*

Key words: genome and metagenome assembly, cluster architecture

Motivation and Aim: *De novo* assembly of large genomes (metagenomes) like the human genome is possible only by using special computer having terabytes of memory, such as SMP-server. Such computers are considerably loose in performance in compare with computing clusters. Standard computing clusters usually have a large number of nodes that makes them an effective tool for solving the time-costly tasks. However, a distributed memory architecture they have is not supported by existing software tools for genome assembly. The aim of the work was to develop a method of assembling genomes from high performance sequencing data with the use of computer clusters with distributed memory.

Methods and Algorithms: The algorithm for assembling of short reads was based on graph overlap approach. Parallelization was performed using MPI technology.

Results: A method for assembly of large genomes on computer clusters with distributed memory was created. It was shown that the method is comparable in quality with existing methods of genome assembly and requires less processing time.

Conclusion: Cluster architecture can significantly accelerate the process of *de novo* assembling large genomes. Our method can be widely implemented for fast genome and metagenome assembly with the use of cluster architecture.

Acknowledgements: This work was financially supported by project VI.61.1.2 to V.A.I. and supported by the Grants Council (under RF President) for State Aid of Leading Scientific Schools (grant Nsh-860.2014.1) to P.S.D.

PROBABILISTIC FRAMEWORK FOR GENE-TREE RECONSTRUCTION AND RECONCILIATION TO A KNOWN SPECIES TREE IN THE PRESENCE OF DUPLICATION, LOSS AND HORIZONTAL GENE TRANSFER

Diallo A.B.*
Dept. Computer sciences, University of Quebec at Montreal, Canada
e-mail: diallo.abdoulaye@uqam.ca

Key words: gene tree reconstruction, horizontal gene transfer, microbe evolution, macroevolutionary processes

Motivation and aims: With the availability of thousands of gene families, new methods for reconstructing gene trees arising from the macroevolutionary processes of duplication, loss and horizontal gene transfer are needed. One way to integrate these events in gene tree construction is with an explicit macroevolutionary model (sometimes referred to as gene tree-species tree reconciliation) [1]. The problem is difficult and previous efforts to include all kinds of events have mitigated results; or can be applied only to a small number of taxa [2] or to parsimony-based approaches [3]. Another important issue is the uncertainties attached to gene tree.

Methods and algorithms: Here we introduce a probabilistic framework for the gene tree reconstruction and reconciliation against a known species tree while accounting for gene tree uncertainty by integrating reconciliation into the tree-building process. he uncertainty of the gene tree is taken into account by allowing a set of bootstrap trees obtained from regular maximum likelihood tree inference methods. For the reconciliation procedure, we divided the species tree in a set of epochs where each epoch contains only co-existing lineages and we modelled each type of events as a Poisson process in the different epochs of speciation. Then the main reconciliation procedure is done in two steps, a reconciliation seed can be quickly computed through a dynamic programming method while mapping all the gene nodes into the species nodes. This quick heuristic step is followed by a hill climbing method to explore the reconciliation space.

Results: The evolutionary histories of ten thousand gene trees were simulated along a reference Tree of Life with different rates of duplication, loss and horizontal gene transfer. The preliminary results clearly indicate that trees constructed taking into account macroevolutionary processes and gene tree uncertainty are more accurate than purely sequence based methods, with RF-distance differences of about 10% on average.

Conclusion: It is worth noting that this method can be easily applied to hundreds of taxa and explores a large reconciliation space. Thus this framework may be useful for the reconstruction of microbe evolution and the study of host and parasite co-evolution.

References:

1. R. Page (1994) Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. *Systematic Biology*, 43(1): 58-77.
2. L. Arvestad et al. (2004). Gene tree reconstruction and orthology analysis based on an integrated model for duplications and sequence evolution. *RECOMB*: 326-335.
3. M. Bansal et al. (2012) Efficient algorithms for the reconciliation problem with gene duplication, horizontal transfer and loss. *Bioinformatics* 28(12):i283-i291

A BIOPHYSICAL APPROACH TO BACTERIAL TRANSCRIPTION START SITE PREDICTION

Djordjevic M.
Institute of Physiology and Biochemistry, University of Belgrade, Faculty of Biology, Belgrade, Serbia
e-mail: dmarko@bio.bg.ac.rs

Key words: transcription start site predictions, bacterial promoters, transcription regulation, biophysical modeling, transcription initiation

Motivation and Aim: Transcription start sites (TSS) in bacterial genomes are locations where RNA polymerase binds and initiates transcription. Accurate knowledge of TSS is important not only for bioinformatic applications (e.g. non-coding RNA gene and operon predictions), but also as the first and the rate limiting step in understanding transcription regulation. TSS predictions in bacteria is a classical bioinformatics problem, where available methods show poor accuracy.

Methods and Algorithms: We approach the problem of TSS detection from a biophysics perspective, where the main idea behind our apprach is to combine accurate alignments of promoter elements [1] with a biophysical model of transcription initiation that we recently developed [2]. The model takes into account both binding of RNA polymerase to double-staranded DNA, and subsequent opening of the two DNA strands, which allows calculating parameters of transcription initiation on a genome-wide scale [2].

Results: We will discuss both theoretical modeling of transcription initiation, and our recent advances in understanding promoter specificity. Improved description of promoter specificity in itself leads to a significant (~50%) reduction in the number of false positives [3], and ability to accurately predict TSS in (relatively short) bacteriophage genomes [4]. We will also present how the modeling and the analyzed sequence specificity are combined in a biophysics based algorithm for TSS detection, which further significantly improves the search accuracy.

Conclusion: We here show that accurate TSS detection requires a change of paradigm, where purely statistical approaches (e.g. inforomation-theory, neural networks, support vector machines) are substituted by course-grained modeling of transcription initiation and a physical understanding of RNAP interactions with promoter sequences. We also show how these ingradients can be integrated in a novel method for TSS detection that significantly improves the search accuracy.

Acknowledgement: This work is supported by a Marie Curie International Reintegration Grant within the 7th European community Framework Programme (PIRG08-GA-2010-276996) and by the Ministry of Education and Science of the Republic of Serbia under project number ON173052.

References:

1. Djordjevic M., J Bacteriol. 193:6305, 2012.
2. Djordjevic M., Integr. Biol. 5:796, 2013.
3. Djordjevic M., to be submitted, 2014.
4. Pavlova O. et al., J. Mol. Biol. 416: 389, 2012.

MODELING BACTERIAL IMMUNE SYSTEMS: CRISPR/CAS REGULATION

Djordjevic M.[1]*, Severinov K.[2, 3], Djordjevic M.[4]
[1]*Institute of Physics, University of Belgrade, Belgrade, Serbia;*
[2] *Institutes of Molecular Genetics and Gene Biology RAS, Moscow, Russia;*
[3]*Waksman Institute, Rutgers, State University of New Jersey, Piscataway, USA;*
[4]*Faculty of Biology, University of Belgrade, Belgrade, Serbia*
e-mail: magda@ipb.ac.rs

Key words: CRISPR/Cas, immune system modeling, synthetic gene circuits, small RNAs

Motivation and Aim: CRISPR/Cas (Clustered Regularly Interspaced Short Palindromic Repeats/CRISPR associated sequences) is a recently discovered prokaryotic immune system [1]. CRISPR cassette is transcribed as a continuous transcript (pre-crRNA), which is processed by Cas proteins into small RNA molecules (crRNAs) that are responsible for defense against invading viruses [2]. Recent experiments show that, surprisingly, overexpression of cas genes generates a large number of crRNAs, from only few pre-crRNAs [3]. This result also suggests a possibility for constructing synthetic gene circuits, which allow generating large product amounts from small amounts of potentially toxic substrate.

Computational methods: We use deterministic and stochastic modeling to analyze CRISPR/Cas system regulation. Inspired by CRISPR/Cas system, we also investigate design of synthetic gene circuit that can generate a large amount of useful product from small amounts of potentially toxic substrate [5]. We optimize the circuit, with the goal of generating maximal product amounts, without increase of substrate amounts upon system induction.

Results: We show that CRISPR/Cas system acts as a strong linear amplifier upon cas overexpression [4]. Interestingly, this strong amplification crucially depends on fast non-specific degradation of pre-crRNA by an unidentified nuclease, which suggests that this nuclease is a major control element of CRISPR/Cas response. We also show that overexpression of cas genes above a certain level does not result in a further increase of crRNA, but that this saturation can be relieved if the rate of CRISPR transcription is increased. Consequently, joint regulation of the system at transcription and post-transcription (transcript processing) level allows rapidly generating thousands of crRNA molecules in a small time-frame of opportunity between the virus infection and the cell lysis. Our results show that such optimal system can rapidly increase the product amount for up to three orders of magnitude, while keeping the substrate amount at constant level.

Conclusion: We analyzed a gene circuit that is capable of rapidly generating a large product amount (crRNAs), from small amount of substrate (unprocessed transcripts). Surprisingly, this rapid generation crucially depends on fast non-specific degradation of substrate molecules. The rapid product generation can be further enhanced by control at the level of transcription initiation.

References:

1. Makarova K.S., et al., Biology Direct 1:7 (2006).
2. Barrangou R., et al., Science 315, 1709 (2007).
3. Pougach K., et al., Mol. Microbiol. 77, 1367 (2010).
4. Djordjevic M., Djrodjevic M., Severinov K. Biology Direct, 7:24 (2012).
5. Djordjevic M., Djordjevic M., Phys Biol. 9(5):056004 (2012).
6. Djordjevic M., Djordjevic M., to be submitter (2014).

GENETIC DISSECTION OF THE INFLORESCENCE BRANCHING TRAIT IN DIPLOID, TETRAPLOID AND HEXAPLOID WHEATS

Dobrovolskaya O.[1]*, Amagai Y.[2], Pont C.[3], Martinek P.[4], Krasnikov A.A.[5], Orlov Y.L.[1], Salina E.A. [1], Salse J.[5], Watanabe N.[2]

[1]*Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia;*
[2]*College of Agriculture, Ibaraki University, 3-21-1 Chuuo, Ami, Inashiki, Ibaraki 300-0393, Japan;*
[3]*INRA-UBP UMR-1095, 5 Chemin de Beaulieu, 63100 Clermont –Ferrand, Cedex 2, France;*
[4]*Agrotest Fyto, Ltd., Havlíčkova 2787, 767 01 Kroměříž, Czech Republic;*
[5]*Central Siberian Botanical Garden SB RAS, Novosibirsk, Russia*
e-mail: oxanad@bionet.nsc.ru
** Corresponding author*

Key words: inflorescence architecture, inflorescence development, independent mutants, SEM, molecular genetic mapping, synteny

Motivation and Aim: Wheat inflorescence, or spike, is characteristically unbranched and normally bears one spikelet per rachis node. Wheat mutants associated with the development of branched spike or multirow spike (MRS) or supernumerary spikelets (SS) constitute key resources in understanding the genetic mechanisms underlying wheat inflorescence architecture and, ultimately, yield components. Here, we report the characterization of genetically unrelated wheat mutants that allowed determine precise positions of the bh/SS genes on molecular-genetic maps, define the candidate genes for the branching/SS/MRS trait and detect developmental defects caused by the bh/SS/mrs1 genes.

Methods and Algorithms: Our strategy integrates analysis of developing inflorescences of wheat mutants using light microscopy and scanning electron microscopy (SEM), molecular genetic mapping with SSR and COS (Conserved Orthologous Set) markers and a candidate gene approach.

Results: The anatomy of young inflorescence of mutant and wild-type lines was compared using light microscopy and SEM. We detected similar abnormal development during early stages of inflorescence development in diploid, tetraploid and hexaploid mutants: development of ectopic spikelets at the location of florets, which always developed at 90° to normal spikelets, or ectopic spikelet of previous order, suggesting that the defect on the mutant occurs after the acquisition of spikelet meristem identity. Molecular genetic mapping placed the genes for the branching/SS trait on the group-2 homoeologous chromosomes (2AS of diploid and tetraploid wheats, and 2AS and 2DS of hexaploid wheat) in the regions of shared synteny. Although the 2AS gene is also involved in control of the SS trait in hexaploid wheat, the 2DS gene is considered as the main contributor at the hexaploid level. Accurate genetic maps that included the *mrs1* gene were used for further synteny-based identification of SS candidate markers using the COSs available in bread wheat. The COS2-COS3 region shared synteny with the rice chromosome 7, *Brachypodium* chromosome 1, and sorghum chromosome 2 regions.

Conclusion: Thus, the genes for the branching inflorescence/SS/MRS trait in diploid, tetraploid and hexaploid wheats are located in the regions of shared synteny of the group-2 homoeologous chromosomes. The genes cause similar abnormalities in the inflorescence development. The mutant phenotypes suggest that the genes are involved in the determining spikelet meristem fate.

Acknowledgements: This work was supported by RFBR grants 12-04-00897-a and 13-04-90932.

ANALYSIS OF BREAD WHEAT LEAF PUBESCENCE FORMATION AND DIVERSITY UZING IMAGE ANALYSIS TECHNIQUE AND MATHEMATICAL MODELING

Doroshkov A.V.*, Zubairova U.S., Genaev M.A., Nikolaev S.V., Pshenichnikova T.A., Afonnikov D.A.
Institut of Cytology and Genetics SB RAS, Russia
e-mail: ad@bionet.nsc.ru
**Corresponding author*

Key words: trichome development, bread wheat, high throughput phenotyping, mathematical model

Motivation and Aim: Plant leaf pubescence (hairiness) plays an important biological role in protection against insect and pathogen attack as well increasing tolerance to abiotic stress conditions. In bread wheat leaf hairiness is formed by unicellular trichomes and displays wide phenotypic variation. However, this trait until recently has been methodologically difficult to phenotype. Currently cereal trichome genetic control and developmental mechanisms is poorly understood. We apply systems biology approach to disclosure mechanisms of pubescence formation in wheat.

Methods and Algorithms: In this study, we used text mining technology for the association network reconstruction. Mathematical modeling of leaf epidermis growth and trichome development was was created using Mathemathica 9 package. Leaf hairiness morphology analysis was performed using high throughput phenotyping method (Genaev et. al., 2012; http://wheatdb.org/lhdetect2)

Results: We reconstructed the network of interactions between known leaf pubescence genes using A. thaliana as model organism. This network consists of a several connected fragments responsible for trichome patterning, trichome cell growth and differentiation, and response to stress factors.

Based on the coarse-grained network structure and information about cereal leaf formation a mathematical model of leaf epidermis growth and trichome development was created. The model shows that the reaction-diffusion processes similar to those described *A. thaliana* epidermis combined with of the cereal epidermal growth process features enough to form the observed trichome spacing pattern in wheat. Also considered the processes of trichome cell growth related to trichome size. For model parameters verification and comparing them with the quantitative characteristics of leaf pubescence in wheat was performed a large-scale analysis of this trait using previously developed image analysis method (http://wheatdb.org/lhdetect2). We analyzed a leaf hairiness in genetic collection of hexaploid, tetraploid and diploid wheat forms. Distribution of trichome length and density argued for partial independence of genetic control of patterning and growth of trichomes. Several distinct morphological types of leaf pubescence were identified. The diversity in pubescence was observed within these groups also. We demonstrated that the model is suitable for fitting different trichome distribution patterns related to several morphological types of leaf pubescence. To investigate possible functional role of several genes involved in the control of the leaf pubescence in wheat genetics experiments were performed. This allowed us to describe the role of these genes and interactions between it were in terms of patterning and growth of trichomes.

Conclusion: The analysis shows the effectiveness of the integration of computer and experimental approaches in the study of such traits as leaf pubescence.

References:

1. Genaev M.A. et al. Extraction of quantitative characteristics describing wheat leaf pubescence with a novel image-processing technique //Planta. – 2012. – Т. 236. – №. 6. – С. 1943-1954.

PRIONIZATION OF THE SFP1 PROTEIN IN YEAST DOES NOT MIMIC ITS INACTIVATION AT WHOLE TRANSCRIPTOME LEVEL

Drozdova P.* , Mironova L.
Dpt. of Genetics and Biotechnology, St. Petersburg State University, St. Petersburg, Russia
e-mail: p.drozdova@bio.spbu.ru
**Corresponding author*

Key words: Saccharomyces cerevisiae, yeast, prions, microarray, hierarchical clustering

Motivation and Aim: In *Saccharomyces cerevisiae*, ribosome biogenesis is controlled by several transcription factors including Sfp1p. Earlier, we found that Sfp1p can form a prion [1]. We have applied microarrays to profile gene expression in the prion-containing strain, which has not been reported before, and in the strain deleted for *SFP1*. We aimed to find out (i) how Sfp1p prionization and its absence alter gene expression in our strains and (ii) how these data correspond to Sfp1p absence or overproduction effects reported in literature [2–5].

Methods and Algorithms: Yeast expression microarrays (Agilent) were used. Further analysis was performed with R/Bioconductor packages [5,6] including limma [7], WGCNA [8], gplots [9], and others.

Results: We detected many differentially expressed genes responding to the *SFP1* deletion, and they represent cellular functions controlled by Sfp1p. Sfp1p prionization affected less genes but they also belong to distinct functional groups. Intersections of gene lists show that only some genes are activated or repressed in both conditions. Hierarchical clustering shows that *SFP1* deletion profiles in different strains cluster together while prionization profile clusters with moderate *SFP1* overexpression, likely because of its lesser effect on transcription.

Conclusion: Sfp1p prionization alters gene expression in a way distinct from its absence. Consequences of *SFP1* deletion in a *S. cerevisiae* strain of the Peterhoff genetic collection are consistent with those reported in other strains. We also validate at least limited use of hierarchical clustering to compare results of biologically, but not technically, related microarray studies.

Acknowledgements: We value support of RFBR (14-04-31265) and SPbU (1.50.2218.2013 and 0.37.696.2013).

References:

1. T. Rogoza et al. (2010) PNAS USA, 107: 10573–10577.
2. G. Chua et al. (2006) PNAS USA, 103: 12045–12050.
3. P. Jorgensen et al. (2004) Genes Dev, 18: 2491–2505.
4. Z. Hu et al. (2007) Nat. Genet., 39: 683–687.
5. J. Reimand J. et al. (2010) NAR, 38: 4768–4777.
6. R. Core Team. (2013) Vienna, Austria.
7. R.C. Gentleman et al. (2004) Genome Biol., 5: R80.
8. G.K. Smyth (2005) In: Bioinforma. Comput. Biol. Solut. Using {R} Bioconductor, R. Gentleman et al. (Eds), 397–420 (Springer).
9. P. Langfelder, S. Horvath (2008), 9: 559.
10. G.R. Warnes et al. (2013).

MECHANISM OF MIRNA ACTION DEFINES THE DYNAMICAL BEHAVIOR OF miRNA-MEDIATED FEED-FORWARD LOOPS

Duk M.A.[2]*, Samsonov A.M.[1, 2], Samsonova M.G.[1]
[1]*St. Petersburg State Polytechnical University, St. Petersburg, Russia*
[2]*A.F. Ioffe Physico-technical Institute of the RAS, St. Petersburg, Russia*
e-mail: duk@mail.ioffe.ru
**Corresponding author*

Key words: feed-forward loops, miRNA, noise buffering

Motivation and Aim: Despite of being extensively studied in the past decade, the molecular mechanisms of miRNA action, as well as its role in the cell are still not clear. There are experimental evidences that miRNA regulates gene expression through translational repression and/or mRNA decay, however the contribution and timing of these mechanisms remain unclear. It is general believed that miRNA can buffer the consequences of noise in gene expression, however up to date there are few experimental observations that support this hypothesis.

Here we perform the theoretical analysis of a gene network sub-system, composed of a feed-forward loop (FFL), in which the upstream transcription factor (TF) regulates the target gene via two parallel pathways: directly, and via interaction with miRNA. As the molecular mechanisms of miRNA action are not clear so far, we elaborate three mathematical models, in which miRNA either represses translation of its target, or promotes target degradation, or is not re-used, but degrades along with target mRNA.

Methods and Algorithms: We examined the FFL dynamics quantitatively and at the whole time interval of cell cycle. The following three statements distinguish our study: we obtained exact solutions to the coupled non-linear differential equations, describing the temporal behavior of the loop components given in [1] in some assumptions; rigorously proved the uniqueness of the solutions obtained and found those parameter intervals, for which the noise buffering by FFLs is possible.

Results: We have shown that different mechanisms of miRNA action lead to a variety of types of dynamical behavior of FFLs. In particular, we found that the ability of FFL to dampen fluctuations introduced by TF is the model and parameter dependent feature. We showed, that in general, an action of any noise, having relatively small amplitude, depends on the Hill function type and can be evaluated via a simple formal analysis of the function.

We proposed that the analysis of the temporal behavior of the FFL together with the ability to find out a unique solution for every set of parameters may provide the way to select the most probable mechanism of miRNA action if the type of FFL is known.

Availability: on request from the authors

Acknowledgements: the support of the RFBR grants 14-04-01522 and 14-01-00334 is gratefully acknowledged.

References:

1. Osella M., Bosia C., Cora D., Caselle M. (2010). The Role of Incoherent MicroRNA-Mediated Feedforward Loops in Noise Buffering, *PLoS Computational Biology*, 7, 3, e1001101.
2. M.A. Duk, A.M. Samsonov, M.G. Samsonova (2013). Early stage dynamics modeling of the feed-forward loops with miRNA. Proceedings of the International Conference "Days On Diffraction" pp. 42-46.

LENS CATARACT: EFFECTS OF CRYSTALLINS MODIFICATION

Duzhak T.*, Tsentalovich Y.
International Tomography Center SB RAS, Novosisibirsk, Russia
e-mail: tgduzhak@mail.ru
**Corresponding author*

Motivation and aim: The composition and architecture of the eye lens, with fiber cells filled with stable long-lived crystallins, allows the lens to focus light on the retina. Three major structural proteins known as α-, β-, and γ-crystallins, by virtue of high concentration and their special structural interactions, provide the transparency of the lens. With aging, crystallins accumulate numerous post-translational modifications (PTMs) and undergo aggregation, cross-linking and water insolubilization. These processes lead to the development of the opacity and are a likely cause of age-related cataract. Presently, the sequence of events that lead to the development of the lens opacity is not well understood. However, evidence suggests that a variety of PTMs in crystallins lead to their cross-linking, aggregation and water insolubilization. Because several of the PTMs in lens proteins simultaneously occur during aging, a combination of cataract-initiating changes in crystallins remains unclear. To understand the cataractogenic process, lens researches have attempted to determine the variety and character of PTMs involved in aggregation and water insolubilization of crystallins.

The purpose of this study was characterization of α-, β-, and γ-crystallins representing soluble (WS) and water insoluble (WIS) protein fractions from the cataractous human lens to identify cataract-specific protein PTMs.

Methods: Full-sized and partially truncated proteins from water soluble fraction of cataractous lens were compared with insoluble/urea denaturated fraction of the same lens by 1-D and 2-D gel electrophoresis. Crystallin molecular weights, pattern and PTMs were identified by MALDI TOF/TOF mass spectrometry. FT ICR mass spectrometry analysis was employed for determination of the total compositions of WS and WIS fractions of cataractous lens. Proteomic identification algorithms Scaffold, Mascot and X!Tandem were used to verify non-enzymatic PTMs in crystallins.

Conclusion: The results show that crystallins, especially α-crystallin group, aggregate irreversibly during cataract development. Some aggregate-forming crystallin peptides (from αA- and αB-crystallins) and non-enzymatic PTMs may be involved in initiating and progression of cataract. The obtained data suggest that α-, β-crystallins represent susceptible sites for spontaneous breakdown in long-living proteins following cross-linking aggregation.

Acknowledgments: This research was supported by RFBR (Projects 14-03-00027, 14-03-00453 and 14-03-31189), by the Ministry of Education and Science RF, and by the Division of Chemistry of RAS. We thank our colleagues Nadezhda Galeva and Todd Williams from The University of Kansas (USA) for their assistance in FT ICR and their insights and expertise that greatly assisted the research.

APPLICATION OF MULTILOCI GENOME SCANNING FOR IDENTIFICATION OF THE KARACHAY HORSE GENETIC STRUCTURE

Erkenov T.A.*, Glazko V.I.
Russian State Agrarian University – Moscow Timiryazev Agricultural Academy
e-mail: erkenov_timur@yahoo.com
**Corresponding author*

Key words: multiloci genome scanning, ISSR-PCR, IRAP-PCR, Karachay horse breed

Motivation and aim: Today we know several local horse breeds which are unique in their adaptation ability to certain ecological and geographical conditions where they are reproduced. To manage genetic resources of these breeds we need easy-to-use genetic elements for multiloci genotyping and gene pool control. A special importance in this case belongs to molecular markers of polymorphism of different genomic elements, for example if there is necessity to support such breeds as Karachay which has its own place in Russian horse breeding. This breed has a number of valuable features which make it unique. The main advantage is high workability and endurance in extreme conditions of highland.

Methods and algorithms: A genomic scanning of Karachay breed, Altai breed and horses of trotter breeds (Orlov trotter, American standardbred, Russian trotter and their crosses with American standardbred) was carried out by using of di- and trinucleotide microsatellite primers in PCR: $(AG)_9C$; $(GA)_9C$, $(GAG)_6C$ and $(CTC)_6C$. Each amplicon was considered as a single locus. We estimated the share of polymorphic loci and Polymorphic Information Content (PIC index) of certain loci and averaged for primer. According to M. Nei (DN, 1972) genetic distances were estimated which were used to make a dendrogram by using TFPGA software.

Results: In order to select the most convenient variants of multiloci genotyping of Karachay horses in this paper we carried out a comparative analysis of multiloci spectra of ISSR (Inter-Simple Sequence Repeat) and IRAP (Inter-Retrotransposon Amplified Polymorphism) markers in horse groups: Karachay breed, Altai breed and horses of trotter breeds. We used di- and trinucleotide microsatellites as ISSR-PCR primers and a fragment of terminal repeat of endogenous retrovirus LTR SIRE-1 as an IRAP-PCR primer. Each group of animals had its specific features of number of DNA fragments of different length and their combinations. Amplification spectra acquired in PCR differed significantly even by ranges of amplicon length, while in spectra of (AG)9C and (GAG)6C primers they differed also between breeds. Karachay horses had higher ranges of amplicon lengths in (AG)9C and (GAG)6C spectra. In LTR SIRE-1 spectra they produced 9 fragments while other horse groups had only 7. We also observed differences in polymorphic information content index: Karachay breed showed higher values in (AG)9C spectra and lower values in (GA)9C spectra.

Conclusions: We collected data about unique genomic distribution of inverted repeats of AG microsatellite in Karachay horse breed which made it different compared to other breeds. That fact allows recommendation of ISSR-PCR markers acquired by using of (AG)9C primer for deep investigation of genetic structure of Karachay horse breed and control of its dynamics. Each horse group had its own specific features of amplification spectra, produced by using primers mentioned above.

Availability: A complex of molecular markers used in this study may be successfully applied in selection of Karachay horse breed as well as of other horse breeds.

KINET 1.0 – A NEW WEB DATABASE ON KINETICS DATA AND PARAMETERS FOR *E. COLI* METABOLIC PATHWAYS

Ermak T.V.[1]*, Akberdin I.R.[1], Timonov V.S.[2, 3], Mishchenko E.L.[1], Oshchepkova E.A.[1], Perfilyeva O.A.[1], Smirnova O.G.[1], Khlebodarova T. M.[1], Likhoshvai V.A.[1, 3]
[1]*Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia;*
[2]*Siberian State University of Telecommunications and Information Sciences, Novosibirsk, Russia;*
[3]*Novosibirsk State University, Novosibirsk, Russia*
e-mail: ermak@biomodelsgroup.ru
**Corresponding author*

Key words: biochemical reactions, kinetic parameters, *E.coli*, web database

Motivation and Aim: One of the key challenges in the systems biology is a revealing of regulatory mechanisms that determined and orchestrated a dynamical behavior of metabolic pathways. Quantitative analysis of these pathways by kinetic modeling is likely sole approach to solve the problem. At the same time, reconstruction of an adequate model requires reliable kinetic data (stoichiometric and kinetic parameters, concentrations of metabolites and enzymes) for the individual biochemical reactions from which a whole pathway comprises. This leads to development of a database in which such kinetic data will be accumulated.

Methods and Algorithms: Web application of the KiNET database was developed on Java platform using Vaadin (http://vaadin.com/) and Spring (http://www.springsource.org/) frameworks.

Results: KiNET was developed as a curated database to accumulate and present kinetic data of biochemical reactions in *E. coli* cell and their related information extracted from literature to support modellers in understanding complex biochemical networks. All the data are manually curated and annotated by biological experts. In order to facilitate the data integration, we have added links for all entities from the database to external resources like ChEBI, PubChem and UniProt/Swiss-Prot.

To estimate percentage of unique data in KiNET, we have performed analysis by comparison with kinetic data stored in another well-known Sabio-RK database (http://sabio.villa-bosch.de/). Sabio-RK data for *E. coli* were exported using its web services. We have compared references for kinetic data on each biochemical reaction presented in both databases using PubMed identifiers. The analysis has shown that only 5% of KiNET data are presented in Sabio-RK.

Conclusion: We have presented unique kinetic data of biochemical reactions (experiment conditions, kinetic parameter values, key metabolite concentrations, cell enzymes and others) as the web-source. The database has a new web-based user interface. It offers more convenient access to the database compared to the old application. We assume that KiNET data will be helpful in carrying out large-scale *in silico* studies.

Availability: http://kinet.biomodelsgroup.ru/

Acknowledgements: This study was partially supported by the RFBR grant (13-01-00344-a) and Program of the RAS “Molecular and cell biology” (6.6).

SINE-MODELING OF GENOME LOOP STEP

Erokhin I.L.
National Biotechnological Company, LLC, Moscow, Russia
e-mail: i.erokhin@inbox.ru

Key words: genome structure, junk DNA

The structure of genome in multicellular organisms is based on such logic element as "genome loop step". This element functions by receiving control from a similar previous element and transfers control to 3 new destinations: two next similar elements and an element capable of activating a certain subroutine.

This study is aimed at modeling a genome loop step via short SINE intercepted repeats.

The model is based on the hypothesis that initiation of short SINE is possible only, when there is short RNA forming a nucleoprotein complex with one of the common factors of Pol III transcription. Moreover, this short RNA is a transcription product of another SINE, and a specific SINE transcription produces a short RNA, which jointly with Pol III common transcription factors, is capable of activating a new SINE. This hypothesis provides for a mechanism of control transfer between specific SINE elements.

1. "Genome loop step" is represented by three SINE elements. All of them are absolutely identical promoters activated via one and the same short RNA (or identical copies of this RNA). It means that the promoters of three non-coding genes are absolutely identical. At the same time the sites responsible for connecting their products (short RNA) with other SINE have their distinctions. Due to these distinctions, short RNA emerging as a result of the first SINE transcription is capable of activating the next "genome loop step" (or the next three SINE elements and nothing except them). Short RNA emerging as a result of the second SINE transcription activates the second "genome loop step" (or another three SINE elements and nothing except them). Short RNA emerging as a result of the third SINE transcription activates a certain subroutine (a non-coding gene serving as a basis for a certain network, which includes both coding and non-coding genes activated in a certain order). ID, B1, B2 is an example of such SINE.

2. "Genome loop step" is represented by a pair of SINE elements having a complicated dimeric structure, (for instance, it can be represented by ID and B1 elements in rodents or Alu repeats in primates). The peculiarity of these dimeric structures is that after post-transcription modification as part of nucleoproteins the non-coding RNA produced as result of their transcription is capable of activating two SINE elements with various promoters instead of one. The first non-coding gene in this pair is capable of activating two new "genome loop steps", while the second non-coding gene in this pair is capable of activating a subroutine. At the same time the second non-coding gene in this structure should not necessarily have a dimeric structure.

3. "Genome loop step" is represented by a non-coding gene of complicated trimeric structure. The activation of such gene can immediately perform the functions of a "genome loop step".

This model was created by the author.

L-MOLKERN SOFTWARE ALLOWING FOR POLARIZATION EFFECTS IN FREE ENERGY CALCULATION

Fomin E.S.*, Alemasov N.A.
Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
e-mail: fomin@bionet.nsc.ru
**Corresponding author*

Key words: molecular dynamics, λ-dynamics, polarization, free energy calculations

Motivation and Aim: Calculation of thermal stability is actually the computation of free energy differences between the

wild type protein and its mutant variant in denatured and native states. But mutations involving amino acids with nonzero charges calculated with the λ -dynamics method [1] are generally less accurate than those involving neutral amino acids .

Methods and Algorithms: We present a program, L-MOLKERN, for free energy difference calculations of single point mutations of proteins by the "alchemical" λ-dynamics. To include polarization effects and thus to increase the accuracy of calculations for mutations involving charged amino acid residues we have developed a new method, NET-Q, which is based on the equilibration principle of electronegativity.

Results: For two single point mutations of barnase protein (PDBID: 1A2P), one of which is electrostatically neutral (A32G) and the other is charged (R72G), we calculated free energies differences by the L-MOLKERN. For the neutral mutation A32G, the free energy difference obtained is equal to ΔG=37.6 kJ/mol that is consistent, within the accuracy ± 4 kJ/mol of the AMBER force field, with the value of 36.1 kJ/mol obtained in [2]. As for the charged R72G mutation, our result ΔΔG=19.8 kJ/mol is much closer to the experimental value ΔΔGexp=10.45 kJ/mol [3] than that of ΔΔGcalc=57.67 kJ/ mol calculated without polarization effects in [2].

Conclusion: Thus, inclusion of polarization effects in free energy calculations provides a better agreement between calculated and experimental values for single point mutations involving charged amino acid residues. A detailed description of the methods used and calculation parameters was published in our previous paper [4].

References:

1. J.L. Knight, C.L. Brooks (2009) Lambda dynamics free energy simulation methods, *Journal of Computational Chemistry.* **30**: 1692–1700.
2. D. Seeliger, B.L. de Groot (2010) Protein thermostability calculations using alchemical free energy simulations, *Biophysical Journal*, 98, **10**: 2309–2316.
3. M.D. Kumar, K.A. Bava et al. (2006) ProTherm and ProNIT: thermodynamic databases for proteins and protein–nucleic acid interactions, *Nucleic Acids Research*, **34**: D204–D206.
4. E.S. Fomin, N.A. Alemasov (2012) A study of the thermal stability of mutant barnase protein variants with MOLKERN software, *Russian Journal of Genetics: Applied Research*, 2, **6**: 453–461.

GTML FORMAT FOR GENE NETS DATA REPRESENTATION

Galimzyanov A.V.
Institute of Biology of the Ufa Research Centre of the RAS, Ufa, Russia
e-mail: galim@anrb.ru

Key words: structural and functional system qualities, relation model

Motivation and Aim: Mathematical and bioinformatical approaches to the study of the gene networks (GNs) involve the development of special formats for complex data that are the inputs for computer simulation environments. In my work the General Threshold Models Language (GTML) for the static representation of the gene network models is proposed. This format is used to build the dynamical representations of the gene networks and their dynamical general threshold models in the software «SETIES» and «SETIES J-Core».

Methods and Algorithms: In the text GTML file the model data are described by a set of command lines (start with a "–"). When executing each command one elementary operation on assembly of the gene network model in the program environment is made. The format of a command line is «–<*c*1> <*p*1> <*c*2> <*p*2>», where *c*1 – the object type; *p*1 – the object primary key value (the linear and embedded structures); *c*2 – the object parameter (data attribute of the GN relational model); *p*2 – the value of the object parameter. The elements of a command line (field) are divided by a single space. Actually two elements of a command line form the command: <object type> and <object parameter>. The command parameters are: <value of the primary key of the object> and <value of the object parameter>. The command line orders to the program-loader to get access to the type object *c*1 with the identifier *p*1 and to appropriate to its parameter *c*2 the value *p*2. In this first appearance of a pair of «–*c*1 *p*1» in the text leads to the creation of the corresponding object. The types of objects: *gn* – the gene network, *gb* – the genetic block (*g*-block), *rb* – the regulatory bond (*r*-bond), *rm* – the regulatory module (*r*-module), *rs* – the regulatory site (*r*-site). Primary keys of objects consist of identifiers of objects. Five types of identifiers are used: *gnID* – the GN identifier, *gbID* – the *g*-block identifier in the GN, *rbID* – the *r*-bond identifier in the *g*-block, *rmID* – the *r*-module identifier in the *g*-block, *rsID* – the *r*-site identifier in the *r*-module (all – integers). The number of components in the primary key for the GN – 1, for the *g*-block – 2, for the *r*-site – 3, for the *r*-module – 3, for the *r*-site – 4. The identifiers in the composite primary key are delimited by a symbol «:». Among the parameters of objects are structural, functional, quantitative, visual and semantic characteristics of the GN components (of more than 50 parameters).

Results: The developed format is used for the design of a repository of the machine-readable GTML files supporting primary or reproducing assembly of control gene network models in the environment of the software product family «SETIES». In current state descriptions of the models of typical synthetic functional modules and some natural molecular and genetic systems of gene expression control are presented to repositories.

Conclusion: Thus, the GTML format allows to create adequate representations of control gene networks, making them open for modeling on various computing platforms.

Availability: The program complexes «SETIES» and «SETIES J-Core» for the analysis of the molecular-genetic control system dynamics are available on request from the author.

THE mRNA FEATURES IMPORTANT FOR TRANSLATION INITIATION REVEALING USING RIBOSOME PROFILING DATA

Gluschenko O.[2], Sharipov R.N.[2], Kondrakhin Yu.V.[2], Volkova O.A.[1]*
[1]*Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia*
[2]*Design Technological Institute of Digital Techniques SB RAS, Novosibirsk, Russia*
e-mail: ov@bionet.nsc.ru
**Corresponding author*

Motivation and Aim: Regulation of eukaryotic genes expression at the translational level mainly occurs at the initiation stage. It is known that 5'-untranslated region (5'UTR) of mRNA is involved in the interaction with the translation initiation factors and 40S ribosomal subunits. Characteristics of 5'UTR nucleotide sequence have a significant impact on the translation initiation efficiency. However, lack of systematic experimental data does not allow identifying the actual important translation initiation parameters and their contribution to the translation initiation efficiency. Due to a new technology of ribosome profiling (RiboSeq), it became possible to estimate the amount of all translated mRNAs, the number of recognized translation initiation sites and the translation speed. These data can be used for identifying the 5'UTR nucleotide sequence parameters important for translation initiation.

Methods and Algorithms: In this work we used the following BioUML (http://www.biouml.org) capabilities: import/export data in different formats; work with tables and samples; access to the sequences and their annotations by DAS (http://www.biodas.org) protocol; the genome browser for interactive sequences visualization, annotation and work with NGS data; multiple methods for data analysis; integration with R / Bioconductor, integration with Galaxy (https://main.g2.bx.psu.edu/), BWA, Bowtie-primarily to align reads.

Results: We carried out a preliminary analysis of SRA data taken from the article (Ingolia et al, 2011) using a BioUML workflow and compared it with the results by Ingolia and coauthors. Correlation of our and Ingolia's results was ~ 0.98.

We collected SRA data on various ribosome profiling experiments and formed mRNA sequence samples of genes. We extracted the following *Mus musculus, Homo sapiens, Arabidopsis thaliana* and *Saccharomyces cerevisiae* mRNA samples from the Ensembl database: complete mRNAs, 5'UTRs, CDSs, 3'UTRs and first introns. The set of some mRNA features like RNA length, 5'UTR length, CDS length, AUG-codons number, AUG-codons with optimal context number was taken for further.

We plan to reveal the mRNA features important for translation initiation using ribosome profiling data and construct a quantitative model of translation initiation, which will allow to predict the efficiency of mRNA translation based on its nucleotide sequence.

This work was supported by the Russian Foundation for Basic Research (14-04-01284).

COMPUTER ANNOTATION OF BACTERIAL GENES USING PHYLOGENETIC PROFILES

Golyshev M.A.[1], Korotkov E.V.[1, 2]*
[1]*Bioinfomatics laboratory, Centre of Bioengineering RAS, Moscow, Russia;*
[2]*Cybernetics Department, National Research Nuclear University «MEPhI», Moscow, Russia*
**Corresponding author*

Over the last years a great number of bacterial genomes were sequenced. Now one of the most important challenges of computational genomics is the functional annotation of nucleic acid sequences, and especially of nucleic acid sequences of genes. The current mathematical methods of functional annotation require the high level of homology between two sequences to reliably predict the possible gene function. Therefore, a sufficient share of the sequenced genes from bacterial genomes is not yet annotated. In this study we presented the computational method and the annotation system for predicting biological functions using phylogenetic profiles. The phylogenetic profile of a gene was created by way of searching for similarities between the nucleotide sequence of the gene and 1204 reference genomes, with further estimation of the statistical significance of found similarities. The profiles of the genes with known functions were used for prediction of possible functions and functional groups for the new genes. We conducted the functional annotation for genes from 104 bacterial genomes and compared the functions predicted by our system with the already known functions. For the genes that have already been annotated, the known function matched the function we predicted in 63% of the time, and in 86% of the time the known function was found within the top five predicted functions. Besides, our system increased the share of annotated genes by 19%. The developed system may be used as an alternative or complementary system to the current annotation systems and for creating functional groups of genes. The coincidence with functional groups of KEGG data bank is more than 50%. The demo version of system is: http://genefunction.ru.

NOVEL microRNAs PREDICTION IN NON-MODEL ORGANISMS

Gruzdeva N.M.[1]*, Nedoluzhko A.V.[1], Shulga O.A.[3], Prokhortchouk E.B.[2], Skryabin K.G.[1, 2]
[1]NRC Kurchatov Institute;
[2]Centre «Bioengineering» RAS
e-mail: nmg76@mail.ru
**Corresponding author*

Key words: microRNA prediction, non-model organisms, high-throughput sequencing

Motivation and Aim: MicroRNAs (miRNAs) are endogenous, non-coding RNA molecules with different regulatory functions. These small RNA oligonucleotides play an important role in cellular life by repressing gene expression. Recent technological advancements, including high-throughput genome sequencing, enabled genome-wide discovery of novel miRNAs and different bioinformatics tools' development. This in turn allowed deeper investigation of biological processes and regulatory networks.

Methods and Algorithms: Here, the miRDeep2 software package was applied to identify miRNAs from high-throughput sequencing data (SOLiD™ 4 System) of non-model organisms' genomes. Genomes of closely related species were used as a reference.

Results: A total of 10 putative novel miRNAs were predicted. The parameters, such as minimum free energy, stability of secondary structures, excision length, etc., were examined.

Conclusion: This study provides the basis for the following investigation of the biological roles of identified miRNAs.

5’ AND 3’ BREAKPOINTS OF mtDNA DELETIONS SHOW DRASTIC DIFFERENCES IN DINUCLEOTIDE PROPERTIES

Gunbin K.V.[1]*, Popadin K.Y.[2]
[1]*Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia*
[2]*University of Geneva Medical School, Geneva, Switzerland*
e-mail: genkvg@bionet.nsc.ru
**Corresponding author*

Key words: mitochondrial DNA, dinucleotide properties, deletion breakpoints

Motivation and Aim: Damas et al 2012 [1] describes a comprehensive survey of non-B DNA conformations across the complete human mitochondrial genome and provide multiple lines of evidence for their association with mtDNA deletions. Moreover it has been demonstrated that mtDNA deletions distributed very unevenly across mtDNA and non-B DNA conformations explains only a part of breakpoint hotspots. Therefore the reason for such breakpoints inhomogeneity remains to be determined.

Methods and Algorithms: We extracted 16,810 complete mtDNA sequences of modern humans from the PhyloTree.org (30 Sep 2012). Then using SATe program v2.2.7 we aligned mtDNA sequences and reconstructed mtDNA consensus (16569 nt). The analysis of inhomogeneity of dinucleotide properties of mtDNA consensus was based on DiProDB data [2], and methods implemented in R boot library, unique breakpoint locations extracted from supplementary data of [1]. The analysis performed using 4 differed sliding windows (2000, 1000, 500 and 300 nt).

Results and Conclusion: It was found that the majority of breakpoints (>92%) locates in proximity (20 nt or less) to mtDNA regions characterized by significant over- or under-representation of various dinucleotide properties. It is of interest that under-representation of dinucleotide properties belonging to B-DNA class (DiProDB classification) describes ~70% of 5’ breakpoints, but only 42% of 3’ breakpoints. This gap increase when we analyzed the breakpoint hotspots (2 or more breakpoints located at similar position): under-representation of B-DNA class of dinucleotide properties describes >72% of 5’ breakpoints but only 27% of 3’ breakpoints. We also showed that dinucleotide properties mostly associated with 5’ and 3’ breakpoint hotspots are quite different: under-representation of DNA persistence length and/or free energy describes >47% and ~50% of 5’ hotspots, respectively; but the most related properties with 3’ breakpoint hotspots are increase in thymine and/or keto (GT) content, ~60% of hotspots. Taking into account that DNA persistence length and free energy properties describe the rigidity of DNA and very important in protein-DNA interactions, the causes of 5’ breakpoint probably related with anomalies in protein-DNA binding functions. In contrast, poly A regions overrepresented in 3’ breakpoint hotspots usually exhibits a single-stranded helical structure which is easy to break.

References:

1. J. Damas, et al. (2012) Mitochondrial DNA deletions are associated with non-B DNA conformations. Nucleic Acids Res. 40:7606-7621.
2. M. Friedel, et al. (2009) DiProDB: a database for dinucleotide properties. Nucleic Acids Res. 37:D37-D40.

HOMO SAPIENS DENISOVA CRAFTSMANSHIP CAN BE RELATED WITH EVOLUTION OF THE miRNAs REGULATING mRNAs EXPRESSED IN THE BRAIN REGIONS CRUCIAL FOR CONSCIOUSNESS AND SPEECH

Gunbin K.V.[1]*, Afonnikov D.A.[1], Kolchanov N.A., Rogaev E.I.[2], Derevianko A.P.[3]
[1]*Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia;*
[2]*University of Massachusetts Medical School, Worcester, Massachusetts, USA;*
[3]*Institute of Archaeology and Ethnography SB RAS, Novosibirsk, Russia*
**Corresponding author*

Key words: miRNA, human evolution, brain

Motivation and Aim: Taking into account that the evolution of miRNA genes was found to be one of the key factors in the formation of the modern type of human, we carried out the analysis of the evolution of miRNA genes and the expression of their mRNA-targets in *Homo sapiens sapiens* (H.s.s.), *Homo sapiens denisova* (H.s.d.) and *Homo sapiens neanderthalensis* (H.s.n.). Using an improved version of the H.s.d. and H.s.n. genomes [1, 2], we performed computer-assisted comparison between the H.s.s. and H.s.d. and the H.s.s. and H.s.n. genomes in order to reveal structural and functional organization of microRNAs (miRNAs) as well as mRNAs targeted by these miRNAs.

Methods and Algorithms: We selected the best-sequenced H.s.d. and H.s.n. pre-miRNAs that are orthologous to H.s.s. pre-miRNAs (miRBase, rel. 19). Then we revealed all the nucleotide substitutions in the ortholog pairs of pre-miRNA H.s.s./H.s.d. and H.s.s./H.s.n.. We selected the H.s.d. and H.s.n. pre-miRNAs that had (i) nucleotide substitutions (deletions, insertions) in the regions that correspond to sequences of mature miRNAs (and/or miRNAs*) or (ii) multiple densely spaced substitutions. From these pre-miRNA groups, we have excluded the pre-miRNAs that occur in the H.s.s. genome in more than one copy. The functional annotation of selected miRNA genes was performed in three steps. First, on the basis of MiRGator 3.0 and Rogaev lab data we selected human miRNAs that have mutated H.s.d. and H.s.n. orthologs and expressed in the central nervous system. Second, for these miRNAs, their target mRNAs were identified using MiRGator 3.0. Third, a search for human brain structures, in which target mRNAs are expressed at increased levels (target brain structures), was performed.

Results: In both H.s.d. and H.s.n. genomes, we have found miRNA genes with fixed unique substitutions miRNAs. The H.s.d. miRNAs containing unique substitutions have been shown to regulate significantly more ($p<1*10^{-5}$) mRNAs that have the highest levels of expression in the brain cortex and thalamic regions [3] responsible for writing, heard speech, spoken language and reading, comparing to H.s.n. miRNAs with unique substitutions also.

Conclusion: Our result highlights the great importance of the evolutionary changes in the sequences of miRNAs in the course of H.s.d. evolution. The result is consistent with the hypothesis that H.s.d., in addition to the fine technical and typological properties of tools and decorative items reported from excavation sites, had a high level of intellect.

The work was supported by budget project VI.61.1.2, RFBR (13-06-12063) and State contract №14.B25.31.0033.

References:

1. M. Meyer, et al. (2012) Science. 338:222-226.
2. K. Prufer, et al. (2014) Nature. 505:43-49.
3. C. Boeckx, A. Benнtez-Burraco (2014) doi: 10.3389/fpsyg.2014.00282.

MOLECULAR EVOLUTION OF MAMMALIAN ORTHOLOGOUS PROTEIN GROUPS INVOLVED IN STEM CELL SPECIFICITY

Gunbin K.V.*, Afonnikov D.A., Orlov Y.L.
Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
**Corresponding author*

Key words: stem cell specificity, protein evolution

Motivation and Aim: Analysis of embryonic stem cells (ESC) specific transcription factors is important for reprogramming and medical applications. Fundamental problem of evolution of stem cell specific proteins could be studied via otrhologous proteins group. As observed previously [1], there is a relation between rare amino acid fixations and shifts in morpho-physiological evolution occurred at relatively recent paleontological periods. In the current study we analyzed the molecular evolution of Mammalian orthologous protein groups involved in stem cell specificity.

Methods and Algorithms: Orthology relations were taken from OrthoDB v.6 [2]. To select highly related sets of proteins we define compact orthologous protein groups (COPGs) as a subgroups of diffuse orthologous protein groups (OPGs) containing proteins from at least 5 species with monophyletic origin. The ancestral sequence reconstruction was performed using RAXML 7.4.2 on the basis of best protein alignment (AQUA 1.1). The COPG trees were corrected on the basis of well-known species tree (TREEFIX) and pre-calculated amino acid replacement matrices (MODELESTIMATOR 1.1). For each observed well-confirmed amino acid replacement type we compared the observed number of changes with expected ones using approach describing elsewhere [3]. The list of proteins involved in stem cell specificity was taken from [4].

Results: It was shown that 7 (of 14) embryonic stem cell specific proteins containing rare amino acid fixations (FOXD3, GBX2, NR5A2, CRABP2, HCK, GABRB3, TFCP2L1) had fixed rare amino acid substitutions at the Mammalia divergence. Other phylogenetic hotspots of rare amino acid fixations related with Boreoeutheria divergence (FOXD3, TFCP2L1), Afrotheria divergence (KIT, LIFR) and Chiroptera divergence (KIT, FOXD3, EDNRB). It is of interest that COMMD3, and FOXD3 proteins fixed rare amino acid substitutions on sequential tree branches leading to Hylobatidae-Hominidae stem and, Hominidae stem respectively. We showed that embryonic stem cell differentiation/ lineage markers also fixed rare amino acid substitutions. Three protein belonging to this category and fixing vast majority of rare amino acid substitutions are COL1A1 (bone marker), LAMA1 (endoderm marker) and KRT1 (trophoblast marker).

Conclusion: Our result highlights the great importance of rare amino acid substitutions at the Mammalia divergence. The result is consistent with the hypothesis of tight relation between placenta development and evolution of gene networks controlling stem cell specificity.

The work supported by Skolkovo Center for Stem Cell Research (agreement with IOGen RAS #8418-43/2013).

References:

1. K.V. Gunbin, D.A. Afonnikov (2014) This issue.
2. R.M. Waterhouse et al. (2013) Nucl. Acids Res., 41:D358-D365.
3. K.V. Gunbin, A. Ruvinsky (2013) J. Mol. Evol., 76:28-47.
4. J.C. Heng et al. (2010) Cold Spring Harb Symp Quant Biol., 75:237-44.

SELECTIVE SHIFTS IN RECENT EVOLUTION OF METAZOA

Gunbin K.V.*, Afonnikov D.A.
Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
e-mail: genkvg@bionet.nsc.ru
**Corresponding author*

Key words: strict orthologous protein sequences, selective shift, rare amino acid substitutions

Motivation and Aim: As was previously observed [1, 2], there is a relation between rare amino acid substitutions and rapid morpho-physiological evolution on deep inner branches of Metazoa phylogenetic tree. The current study aimed to test the relation between molecular and morpho-physiological evolution on tree branches of the Metazoan phylogenetic tree corresponding to relatively recent paleontological periods.

Methods and Algorithms: The molecular evolution of proteins from 100 Metazoan species was analyzed. Orthology relations between amino acid sequences were taken from OrthoDB 6 [3]. To select highly related sets of proteins we define compact orthologous protein groups (COPGs) as a subgroups of diffuse orthologous protein groups (OPGs) containing proteins from at least 5 species with monophyletic origin. The ancestral sequences reconstruction at each internal tree node of COPG trees was made using RAXML 7.4.2 on the basis of best protein alignment (AQUA 1.1). The COPG trees were corrected on the basis of well-known species tree (TREEFIX) and pre-calculated amino acid replacement matrices (MODELESTIMATOR 1.1). The ancestral sequences were used to estimate the number of observed amino acid substitutions, while the amino acid replacement matrices were used to estimate the number of expected amino acid substitutions under the assumption of a stationary Markov process of protein evolution. For each observed well-confirmed amino acid replacement type we compared the observed number of changes with expected ones using permutation test optimized for high computational performance[1] and determined amino acid replacements highly improbable (atypical) according to Markov model.

Results and Conclusion: For COPGs containing atypical amino acid replacements we conducted the enrichment tests at each inner Metazoa tree node. This allowed to uncover structural (Pfam domains) and functional (GO terms) features overrepresented in COPGs with atypical amino acid replacements. As result, we identified tree branches characterized by drastic changes (in terms of enriched COPG features) in selection force during relatively recent paleontological periods. It is of big interest that the most pronounced changes have been linked to Hominida, Geomyoidea, Euarchontoglires, Formicoids, Culicidae and Drosophilidae divergences. Moreover, it was shown that these divergences characterized by intensive reorganization of protein families involved mainly in morphogenesis and neural system functioning. Thus, it was clearly shown that rare amino acid fixations relate with shifts in morpho-physiological evolution occurred at relatively recent paleontological periods.

The work was supported by “Biosphere origin and evolution” RAS program, budget project VI.61.1.2, MCB RAS project 6.6.

References:

1. K.V. Gunbin, A. Ruvinsky (2013) J. Mol. Evol., 76:28-47.
2. K.V. Gunbin et al. (2011) BMC Evol Biol. 11: 224.
3. R.M. Waterhouse et al. (2013) Nucl. Acids Res., 41:D358-D365.

THE RELATION BETWEEN ENVIRONMENTAL CHANGES AND EVOLUTION OF ARCHAEA PROTEIN DOMAINS

Gunbin K.V.*, Afonnikov D.A.
Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
e-mail: genkvg@bionet.nsc.ru
**Corresponding author*

Key words: Archaea, protein evolution, environmental changes

Motivation and Aim: Given the enormous rate of reproduction, as well as intense horizontal gene transfer, the reconstruction of phylogenetic trees of prokaryotes, especially at the Archean era is difficult. Analysis of protein molecular evolution uncovers the main directions of gene transfer due to the environmental conditions, and the so-called atypical periods associated with intense transformations of habitats. This work was aimed to find such periods of protein molecular evolution in Archaea kingdom.

Methods and Algorithms: We selected 3239 orthologous protein groups (OPG) representing proteins belonging to at least four taxonomic orders of Archaea from OMA orthology database (Dec. 2012). OPG multiple alignment and phylogenetic tree reconstruction were made simultaneously using SATe v2.2.7. The reconstruction of ancestral protein sequences in each inner tree node were made using RAxML v. 7.4.2 on the basis of WAG, LG and JTT models. For each OPG the known protein domains were identified using interproscan-5-RC6. We analyzed Archea-specific protein domains only. Finally, for these domains we estimated evolutionary changes of all physicochemical amino acid properties (PCAAP) available in Aaindex 9.1 database.

Results and Conclusion: We identified 27 core neighborhoods of taxa describing the consensus Archaea tree. The analysis of the PCAAP evolution helped to subdivide core taxa neighborhoods into 2 contrast groups and 2 intermediate ones according to relative rate of changes in various amino acid characteristics. The main class comprising 13 core neighborhoods represents small gradual changes (mainly in hydrophobicity) in Archaea-specific protein domains associated with long-term adaptation of the protein globule to constant environmental conditions. The second class comprising 5 core neighborhoods (Korar.-Nanoar., Korar.-Thaumar., Methanopyr.-Thermoplasm., Nanoar.-Thermoprot., Thaumar.-Thermococ.) represents a drastic change in the core structures (mainly in α-helical propencity) of domains associated with its formation at the dawn of Archaea evolution in the Archean and Proterozoic eras. Two intermediate groups (Korar.-Thermococ., Nanoar.-Thaumar., Thaumar.-Thermoprot., Methanobac.-Methanopyr., Methanococ.-Methanopyr., Halobac.-Thermoplasm.; and Nanoar.-Thermococ., Methanopyr.-Thermococ., Korar.-Sulfolob.) represent an intermediate change (mainly in secondary structure fragmentation) in protein domains associated with severe ecological changes. Comparison of functional annotations (using dcGO) of Archaea-specific protein domains belonging to OPGs representing main class of core neighborhoods with three additional classes allows us to identify biochemical processes evolved. These processes are synthesis of nucleic acids, the formation of macromolecular complexes and energy metabolism. Thus it was clearly shown that all periods of atypical protein evolution of Archaea related with intense transformations of bacteria habitats. Moreover it was shown that the most severe protein and environment changes confined to Archaea tree root.

The work was supported by "Biosphere origin and evolution" RAS program, budget project VI.61.1.2, MCB RAS project 6.6.

WHAT EVOLUTION OF HOMINID TATA-BOXES CAN TELL US ABOUT HUMAN LINEAGE?

Gunbin K.V.[1]*, Ponomarenko M.P.[1], Afonnikov D.A.[1], Gusev F.[2], Rogaev E.I.[2]
[1]*Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia;*
[2]*University of Massachusetts Medical School, Worcester, Massachusetts, USA*
**Corresponding author*

Key words: TATA-boxes, hominids, molecular evolution

Motivation and Aim: The great enigma of contemporary evolutionary biology is how tiny changes of hominid genes could result in the unique human species. In order to answer this question we performed complex evolutionary analysis of core promoter regions of hominid (*H. sapiens, P. troglodytes, G. gorilla, P. pygmaeus*) genes whose human orthologs are expressed in neurons and lymphoid cells.

Methods and Algorithms: Multiple alignments of upstream regions of genes were extracted from the EnsEmbl rel. 73. Human transcription start sites (TSS) were taken from GENCODE rel. 17, validated using RNA-seq peaks data from neurons and lymphoid cells. For each TSS alignment, ancestral nucleotide sequences were reconstructed in each internal node of the Hominidae tree using PAML 4.7. We considered nucleotide positions whose evolution could be reconstructed at a probability higher than 99%. To predict TATA-box activity at each node of the Hominidae tree we implement our algorithm described in [1], which used an empirical equilibrium equation for three-step for TBP/TATA binding: TBP slides along DNA, TBP stops in the TATA box, the TBP/TATA-complex is finally stabilized due to deformations in DNA. mRNAs with biochemically significant changes in TATA-box activity (negative or positive) were chosen for each Hominidae tree branch. Annotation of these mRNAs was performed using expression information from the Allen Human Brain Atlas (March 2013), the Human Protein Atlas v12, the Gene Expression Atlas v.2 and the Gene Expression Barcode v.2. Statistical evaluation of the contribution of mRNAs with evolving TATA-boxes to tissue functioning were made using randomization test.

Results and Conclusion: The differences in nucleotide substitution and/or indel fixation rates in core promoters between different hominid lineages were not detected. Nevertheless, highly unequal rates of TATA-box activity changes in different hominid lineages were found. Significant changes in TATA-boxes activity (both operating in neurons and lymphoid cells) were the most characteristic for lower hominid lineages in contrast with human and hominid ancestor lineages. This fact could be explained by TSS jumping in lower hominids. In the case of TSS jumping the intensive reorganization of core promoters and subsequent change in gene expression regulation must be occurred. Moreover, in vast majority of cases genes regulated by evolving TATA-boxes expressed in various tissues, thus, the evolution of TATA-box activity could leads to change in gene expression not in neurons and lymphoid cells only but in all organism. Taking this into consideration we could conclude the ancestrally/conservatism of human lineage, which expressed in rate of TATA-box activity evolution, in contrast with lower hominid lineages undergoing intensive changing in TATA-box activity.

The work was supported by budget project VI.61.1.2 and State contract №14. B25.31.0033.

References:

1. P.M. Ponomarenko, et al. (2008) A step-by-step model of TBP/TATA box binding allows predicting human hereditary diseases by single nucleotide polymorphism, Dokl Biochem Biophys. 419:88-92.

WHAT EVOLUTION OF RYE SUBTELOMERIC REPEATS CAN TELL US ABOUT CEREALS SPECIATIONS?

Gunbin K.V.[1]*, Levitsky V.G.[1], Vershinin A.V.[2]
[1]*Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia;*
[2]*Institute of Molecular and Cellular Biology SB RAS, Novosibirsk, Russia*
e-mail: genkvg@bionet.nsc.ru
**Corresponding author*

Key words: Secale, pSc200 and pSc250 repeats, sequence conservation, dinucleotide properties

Motivation and Aim: A specific feature of the rye (*Secale cereale*) is large heterochromatin blocks at the ends of all seven pairs of chromosomes. Multicopy tandemly repeated DNA families, pSc200 and pSc250, are among the major components of the rye subtelomeric heterochromatin This work is devoted to understanding the evolution trends of these repeat families and their possible role in the rye hetechromatin formation.

Methods and Algorithms: The Internet available 454 sequencing datasets were exploited: ERX140512 ERX140514 ERX140515 ERX140516 ERX140519 ERX140520, ERX140521. We cut adapter sequences (using TagCleaner 0.12) and selected best reads by PHred quality of reading (sel. condition: $\overline{Q} - 1.6\sigma \geq 10$, where $\overline{Q}$ - read means quality, σ – standard deviation of quality). As a result, it was selected 22626026 reads. From these reads we extracted pSc200 and pSc250 monomers using BLASTN. Then we selected out nearly identical monomers (by MUMmer 3.23), created multiple alignment (using SATe v2.2.7), and tested the robustness of phylogenetic signal by alignment jackknifing (500 iterations, FastTree 2.1.3). On the basis of the pSc200 and pSc250 consensuses resulted from alignments we conducted the analysis of inhomogeneity of dinucleotide properties (DiProDB data, R boot library).

Results and Conclusion: The phylogenetic analysis revealed the star-like phylogeny of pSc250 without distinguishable clusters; in contrast, pSc200 forms four clearly separated clusters. Applied Dendroscope 3.2.10 we reconstructed the phylogenetic networks of the pSc250 and pSc200 families. They contain about 1% of hybrid nodes which recognize the value of the monomers hybridization within families. Detailed analysis of phylogenetic trees shows nearly identical tiny stemness of pSc250 and pSc200 indicating of explosive burst evolution for these repeat during rye speciation. The analysis of dinucleotide conformational properties inhomogeneity along pSc200 and pSc250 consensus shows drastic differences in monomers organization: pSc200 has two long (>15 nt) regions that were characterized by significant high or low values of more than ten DNA conformational properties (such as tilt, rise, major groove width) which nearly perfectly match with evolutionary conservative regions, while pSc250 has only one evolutionary conservative region which did not match with any observed short regions characterized by significant high or low values of not more than three conformational properties. Moreover, pSc200 is widely spreaded among cereals and retains a high level of homology (92-93%), while pSc250 is specific to genus *Secale*. Thus two tamdem repeat families, pSc200 and pSc250, have similar genomic organization and abundance in the rye genome but different evolutionary trends. While pSc200 is most probably involved in some universal molecular mechanism processing in the chromosome behaviour, pSc250 is yet contributed in the hetechromatin enlargement during rye speciation.

IMAGE J ADDON FOR 2D ELECTROPHORESIS GEL ANALYSIS

Gurkov A.N.*, Kondratyeva E.M., Bedulina D.S.
Institute of Biology at Irkutsk State University, Irkutsk, Russia
e-mail: a.n.gurkov@gmail.com
**Corresponding author*

Key words: ImageJ, 2D electrophoresis, 2D gel analysis software

Motivation and Aim: Two-dimensional gel electrophoresis of proteins is widely used in modern proteomics. There is a number of commercial software for analysis of obtained 2D gels, however there is no free software, which provides similar functionality for full and automated analysis. Existing free software like RegStatGel [1] and Pinnacle [2] perform only certain steps of analysis, but not all of them. Whereas ImageJ is a perfect base for this kind of image analysis: it is free, open source and cross-platform, it is very popular and provides powerful extensibility. A workflow for 2D gel analysis with ImageJ have been published [3], but the algorithm contains some mistakes and no macro or plugin have been published to automate the workflow. The aim of this work was to develop an addon for ImageJ, which allows to fulfil full analysis of 2D electrophoresis gels as automatically as possible.

Methods and Algorithms: The addon represents the media that brings together under ImageJ or Fiji several ImageJ plugins, new implementation of developed earlier algorithm and new algorithms to perform following steps of 2D gel analysis: image alignment (plugins «Extract SIFT Correspondences», «Extract MOPS Correspondences» and «bUnwarpJ»), protein spots identification (implementation of foreign algorithm [4], original algorithm or plugin «Watershed Algorithm») and spots quantification (original algorithm).

Results and conclusion: The worked out addon can substitute existing commercial software and provide the improved and highly automated workflow for 2D-PAGE analysis for free.

Availability: Debugging of the addon is finishing now. The addon is available on request from the authors, but soon it will be available through the web-pages of ImageJ and Fiji.

References:

1. Li F., Seillier-Moiseiwitsch F. (2010) Differential analysis of 2D gel images, Methods in enzymology, 487: 595-609.
2. Morris J.S. et al. (2009) Evaluating the performance of new approaches to spot quantification and differential expression in 2-dimensional gel electrophoresis studies, Journal of proteome research, 9, 1: 595-604.
3. Natale M. et al. (2011) Image Analysis Workflow for 2-D Electrophoresis Gels Based on ImageJ, Proteomics Insights, 4: 37-49.
4. Kostopoulou E. et al. (2014) An Effective Approach for Detection and Segmentation of Protein Spots on 2-D Gel Images, Biomedical and Health Informatics, 18, 1: 67-76.

CONTROLLED VOCABULARIES AND INFORMATION TABLES FOR THE KNOWLEDGE BASE ON EPIGENETIC CONTROL OF HUMAN EMBRYONIC STEM CELLS

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

Key words: embryonic stem cells, pluripotency, chromatin-modifications, chromatin-modifying factors, transcriptional regulation, knowledge base, vocabularies

Motivation and Aim: Pluripotency of embryonic stem cells (ESCs) and induced pluripotent stem cells (iPSCs) is maintained by complex system that includes genetic and epigenetic levels. A great variety of regulatory mechanisms and regulatory proteins (transcription factors, mediator proteins and co-regulators including proteins with chromatin-remodeling and chromatin-modifying activities) are involved in transcriptional control of pluripotency. The purpose of this work was to develop controlled vocabularies and information tables for the knowledge base on epigenetic control of gene expression in human ESCs and iPSCs. This knowledge base may be useful for analysis and interpretation of high-throughput transcriptomic, proteomic and metabolomics data as well as data on 3D organization of the human genome, etc.

Methods and Algorithms: Data for vocabularies were extracted from scientific literature and public resources: i) Entrez Gene Database; ii) UniProt; iii) CREMOFAC; iv) CR Cistrome; v) PluriNetWork.

Results: Three hierarchical vocabularies were developed: (i) chromatin-modifications and its effect on transcriptional activity; (ii) chromatin regulators and its functions; (iii) chromatin regulatory complexes. The following information tables for accumulation data on genetic and epigenetic control of gene expression in ESCs and iPSCs were designed: (i) the effects of chromatin modifications on transcriptional activity of the gene or coactivator (corepressor) complexes assembly; (ii) the effects of transcription regulators (transcription factors or chromatin-modifying (-remodeling) factors) on chromatin modifications, transcriptional activity of the gene or coactivator (corepressor) complexes assembly.

Conclusion: The future prospect of this work is developing of the knowledge base on epigenetic control of gene expression in the human ESCs and iPSCs. We believe that such knowledge base may give a deeper view of molecular-genetic basis of epigenetic control of pluripotency and differentiation of ESCs and iPSCs. Hierarchical vocabularies may be useful for of high-throughput data analysis.

Acknowledgements: The work supported by Skolkovo Center for Stem Cell Research (agreement with IOGen RAS #8418-43/2013).

FUNCTIONAL CHARACTERISTICS OF HUMAN GENES CONTAINING LOW LEVEL OF PROMOTER POLYMORPHISM REVEALED FROM THE 1000 GENOMES PROJECT DATASET

Ignatieva E.V.*, Levitsky V.G., Kolchanov N.A.
Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
e-mail: eignat@bionet.nsc.ru
**Corresponding author*

Key words: single nucleotide polymorphism, promoter, 1000 Genomes Project

Motivation and Aim: The goal of this study was to reveal functional characteristics of human genes containing extremely low level of SNP in promoter regions. This knowledge may give a deeper view of genic intolerance to regulatory variation and may be useful for interpretation of personal genomes.

Methods and Algorithms: The whole-genome set of 23,372 human transcripts was designed as it was described previously [1]. The promoter SNP content was determined using the 1000 Genomes Project dataset as the count of SNPs in the 500 bp long region upstream of the annotated TSS. Gene Ontology (GO) enrichment analysis was performed using DAVID. Groups with fold enrichments 1.5 or more and standard significance level $p<0.01$ were considered to be interesting. The set of genes encoding transcription factors (TFs) was compiled using TcoF-DB.

Results: A broad variability of SNP contents in promoters of human genes was revealed. At least six SNPs were found in 1,258 (5.5%) transcripts. Almost a one-fifth (18.5%) of the total number of transcripts do not contain SNP in their 500-bp long upstream regions. Among the genes which promoters do not contain SNPs, several functional groups were overrepresented: (1) genes involved in chromatin organization or (2) regulation of gene-specific transcription; genes controlling (3) cell (neuron) projection morphogenesis; (4) myeloid cell differentiation; (5) erythrocyte differentiation and homeostasis; (6) development of primary sexual characteristics. We had also compared the contents of SNPs in the upstream regions of genes encoding TFs with that for the whole-genome dataset. It was shown that promoters of genes encoding TFs were characterized by decreased genetic variability in comparison to that for the whole genome dataset.

Conclusion: Previously it was found that among genes with high promoter SNP content (six SNPs and more) several functional groups were overrepresented significantly [1]. In this study it was demonstrated that the genes involved in expression regulation and genes important for embryonic development were overrepresented among genes which promoters do not contain SNPs. This observation points on a decreased tolerance of these groups of genes to regulatory genetic variation.

Acknowledgements: This work was partly supported by Presidium of the SB RAS (project No. VI.61.1.2 from nominate VI.61)

References:

1. Ignatieva E.V. et al. (2014) Genetic basis of olfactory cognition: extremely high level of DNA sequence polymorphism in promoter regions of the human olfactory receptor genes revealed using the 1000 Genomes Project dataset, *Front Psychol*, 5:247.

THE KNOWLEDGE BASE ON MOLECULAR GENETICS MECHANISMS CONTROLLING HUMAN LIPID METABOLISM

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

Key words: knowledge base, cholesterol biosynthesis, reverse cholesterol transport, transcriptional regulation, gene network, GWAS

Motivation and Aim: Lipid metabolism (LM) implies biosynthetic pathways, transport, intracellular utilization, degradation, and excretion of lipids including such very physiologically important ones as triglycerides and cholesterol. These processes are controlled by numerous proteins and genes. An interest for the LM system is extremely high due to the facts that the lipid abnormalities are associated with an increased risk for vascular disease, and especially heart attacks and strokes. To support experimental investigation of genetic factors predisposing to high blood cholesterol level, the knowledge base on molecular genetics mechanisms controlling human lipid metabolism (LM-KnB) was developed.

Methods and Algorithms: Data were extracted from scientific publications and public databases. The gene network was reconstructed and visualized using GeneEd [1]

Results: At present LM-KnB includes the following data.

1) The gene network representing molecular genetics mechanisms, which control the intracellular cholesterol level: cholesterol biosynthesis, cholesterol uptake by cells and reverse cholesterol transport. The data is presented in the form of diagram which includes objects (proteins, genes, metabolites, etc.), biochemical reactions and regulatory processes. The key regulators are transcription factors of the SREBP family which function as a part of cholesterol sensor and control expression of genes encoding enzymes of cholesterol biosynthesis, uptake and efflux [2].

2) The compilation of human loci and genetic variants effecting plasma concentrations of lipids (total cholesterol, high- or low-density lipoprotein cholesterol, triglycerides), revealed by genome-wide association analysis or large-scale gene-centric meta-analysis.

Conclusion: In future, we plan further development of the LM-KnB. We believe that LM-KnB will give a deeper view of molecular-genetic basis of human diseases and will be useful for designing new pharmacological approaches for treatment.

Acknowledgements: The work was supported by the Presidium of SB RAS (project N 6.6 from the Program “Molecular and Cellular Biology”).

References:

1. E.A. Ananko et al. (2005) GeneNet in 2005, Nucleic Acids Res., **33**: D425-7.
2. T.I. Merkulova et al. (2013) Transcription Regulatory Codes of Eukaryotic Genomes, *Russian Journal of Genetics*, **49**:29–45.

METHOD TO PREDICT THE PERCENTAGE OF CELL TYPES IN HUMAN BLOOD

Igolkina A.A.*, Samsonova M.G.
Petersburg State Polytechnical University
e-mail: igolkinaanna1@gmail.com
**Corresponding author*

Key words: blood, heterogeneous tissue, cell type percentage, expression deconvolution

Motivation and Aim: Blood is the most investigated heterogeneous tissue. It contains a variety of cell types, of which the major types are Lymphocytes, Monocytes, Granulocytes, Erythrocytes, Megakaryocyte (Lymphocytes and Granulocytes are complex cell groups in turn). Gene expression data from blood genomics studies is widely used in medical diagnosis. Most of these studies are based on the analysis of total peripheral blood mononuclear cells (PBMCs). PBMCs are composed of over a dozen cell types, the proportion of which varies in blood samples from individual people. This variability significantly influences genome-wide gene expression data. The heterogeneity of blood distorts the data, however, it is often discarded due to the lack of data on the composition of the samples. The application of experimental methods to separate or quantify constituents from each sample is time-consuming and does not solve the problem. Therefore, an attractive alternative is to accurately deconvolve gene expression data. Here we develop a method to predictthe percentage of cell types in a blood sample from whole genome gene expression data.

MethodsandAlgorithms: We used a data set of 628 patients with known gene expression levels and percentages of 5 cell types in blood samples. The samples were devided into training set, testing set and set for prediction.

WebuiltandtestedvariouspredictivemodelsbasedonPCA, linearregression(with and without prior knowledge of cell type specific signattures obtained from pure cell types [1-2]) and SVM with different kernel types. To select a gene subset which provides the best prediction of cell proportions we use random sampling, filter methods(correlation, mutual info), wrapped methods (RFE), as well as some prior knowledge in the form of sets of marker genes[3], i.e.

It is noteworthy that both feature selection and predictive methods were constructed for each individual cell type independently. To estimate the performance of different approaches a the Pearson correlation coefficient between estimated and true cell type proportion in data was calculated.

Results: We achieved the best estimation of cell type proportions using filter or wrapped feature selection method and SVM with Gausian kernel. This approach significantly improved the Pearson correlation between true and estimated cell type proportions to approximately 0,8-0.95, that is high enough for further studies.

Conclusion: We have developed a method that can accurately predict the percentage of cell types from whole genome gene expression data in human blood samples. Our approach can be used to predict the percentage of cell types in other tissues.

Availability: The MATLAB script is available on request from the author.

References:

1. Alexander R Abbas et al. “Deconvolution of blood microarray data identifies cellular activation patterns insystemic lupus erythematosus.” In: *PloS one* 4.7 (Jan. 2009).
2. Ting Gong et al. “Optimal deconvolution of transcriptional profiling data using quadratic programming with application to complex clinical blood samples.” In: *PloS one* 6.11 (Jan. 2011).
3. Renaud Gaujoux, “An introduction to gene expression deconvolution and the CellMix package”, 2013.

ANALYSIS OF A TOMATO INTROGRESSION LINE, IL8-3, WITH INCREASED BRIX CONTENT USING THE WHOLE-GENOME SEQUENCE

Ikeda H.[1, 2]*, Kanayama Y.[1]
[1]*Graduate school of Agricultural Science, Tohoku University, Sendai, Japan;*
[2]*Research Fellow of Japan Society for the Promotion of Science*
e-mail: h.ikeda@s.tohoku.ac.jp
**Corresponding author*

Key words: Brix, *Solanum lycopersicum*, *Solanum pennellii*, tomato, whole-genome sequence

Motivation and Aim: Tomato is a major crop and a model plant for studying fruit development, and the sequencing of its genome has been completed (The Tomato Genome Consortium, 2012). In tomato, fruit soluble solids (Brix) is one of the most important determinants of fruit quality, and increasing the sugar content is an important objective of tomato breeding. IL8-3, one of the tomato introgression lines developed from a cross between the wild relative *Solanum pennellii* and the cultivated tomato *S. lycopersicum* M82, was investigated to isolate genes associated with Brix content.

Methods and Algorithms: Plant growth, fruit yield, and sugar metabolism were analyzed to understand the physiological mechanism underlying high Brix content in the IL8-3 fruit. Then, recombinant lines were prepared from F_2 plants, and DNA markers were developed on the basis of the whole-genome sequence of tomato to isolate the gene underlying QTL responsible for Brix content.

Results: The high Brix content in the IL8-3 fruit was because of the high hexose content rather than organic acid or sucrose content and could be associated with the high starch accumulation. We genotyped approximately 1500 F_2 progeny of the cross between M82 and IL8-3 with a set of DNA markers and identified 92 recombinants. The recombinant lines were grown for the measurement of Brix content, and a QTL for Brix content was delimited to a 300-kbp region of chromosome 8. Candidate genes within the region were analyzed by a quantitative real-time PCR method to confirm gene expression.

Conclusion and Availability: The physiological mechanism underlying increased Brix content in the IL8-3 fruit was proposed, and a QTL for Brix content was delimited to a narrow region of chromosome 8. Our study showed that the whole-genome sequence of tomato, which is available in SOL genomics network (http://solgenomics.net/), is useful and may accelerate molecular genetic studies in Solanaceae plants.

References:

1. The Tomato Genome Consortium. (2012) The tomato genome sequence provides insights into fleshy fruit evolution, Nature, **485:** 635-641.

MATHEMATICAL MODELING OF LUNG INFECTION AND ANTIBIOTIC RESISTANCE

Ilin A.[1], Islamov R.[1], Kasenov S.[2], Nurseitov D.[2], Serovajsky S.[3]*
[1]*Scientific Center for Anti-infectious Drugs, Almaty, Kazakhstan;*
[2]*National Open Research Laboratory of Information and Space Technologies of KazNTU after K. Satpaev, Almaty, Kazakhstan;*
[3]*Al-Farabi Kazakh National University, Almaty, Kazakhstan*
e-mail: serovajskys@mail.ru
** Corresponding author*

Key words: antibiotic resistance, reverse resistance, population dynamics, mathematical modeling, differential equations, numerical analysis

Motivation and Aim: At various times it was shown that restricting the use of antibiotics in clinical practice resulting in successful decrease in antibiotic resistance (Barbosa T.M. & Levy S.B. 2000). Some authors have attributed this phenomenon to the fitness cost (Andersson D.I. & Hughes D. 2010). However, most researchers pay attention to the development of resistance and transmission. Various models considered these processes in order to develop new treatment strategies (D'Agata E.M.C. et al. 2007). The aim of this work was to develop a model of the observed reversal of antibiotic resistance in therapy of pulmonary infections.

Methods and Algorithms: The process considered of treatment of the infected organism. At its initial stage, the bacteria infected of organism. Further starting treatment organism of some antibiotic. Consequently, the number of bacteria is reduced, and a complete recovery may occur. However, due to mutations appear bacteria resistant to the action of this antibiotic. Over time the effectiveness of the antibiotic is substantially reduced. Treatment continued for a new drug. After some time, the sensitivity of bacteria to the initial antibiotic is gradually recovering. The mathematical model of the observable process is proposed. This is a system of nonlinear differential equations. The model characterizes the variation of the general bacterial population, and as well as mutants resistant to antibiotics. The differential equations theory is applied for the qualitative analysis of this model. The numerical solution of the system is found for the large enough scope of process parameters. The experimental data is used for the assessment of the model efficiency.

Results: The general stages of the observable process are determined by means of the approximate qualitative analysis of the suggested mathematical model. These results are corroborated by computer analysis. The clinical material prepared for validation of the developed model.

Conclusion: By dint of the proposed model is used to simulate the reversal of antibiotic resistance in pulmonary infectious diseases. This will not only describe the process of reversion of antibiotic resistance, but also to develop a new treatment strategy.

References:

1. T.M. Barbosa, S.B. Levy (2000) The impact of antibiotic use on resistance development and persistence, *Drug Resistance Updates*, **3:** 303–311.
2. D.I. Andersson, D. Hughes (2010) Antibiotic resistance and its cost: is it possible to reverse resistance?, *Nature Reviews Microbiology*, doi:10.1038/nrmicro2319.
3. E.M.C. D'Agata et al. (2007) Modeling antibiotic resistance in hospitals: The impact of minimizing treatment duration, *Journal of Theoretical Biology*, **249:** 487–499

FEATURES OF INTERACTIONS BETWEEN miR-1273 FAMILY AND mRNA OF TARGET GENES

Ivashchenko A.T.*, Berillo O.A., Pyrkova A.Y., Niyazova R.E.
National Nanotechnology Laboratory, *Al-Farabi KazNU, Almaty, Kazakhstan*
e-mail: a_ivashchenko@mail.ru
**Corresponding author*

Key words: miRNA, mRNA, miR-1273, binding site, human

Motivation and Aim: miRNAs participate in the post-transcription regulation of gene expression. miRNAs participate in many pathological processes. Changes in the miRNA concentration have been shown to occur during the development of breast, lung, esophageal, stomach, intestine, prostate and other cancers. In this work, we studied the binding of 2,563 miRNAs with 12,175 mRNAs. The majority of these genes participate in the development of cancer and other diseases.

Methods and Algorithms: MirTarget program defines the localization of miRNA binding sites in the 5'UTRs, CDSs and 3'UTRs of the mRNAs; it calculates the free energy of hybridization (ΔG, kJ/mole) and the ratio $\Delta G/\Delta G_m$ (%), where ΔG_m equal to the free energy of miRNA binding with completely complementary nucleotide sequence. The binding sites of miRNAs with mRNAs were selected by $\Delta G/\Delta G_m$ ratio of 90% or more.

Results: The miR-1273 family consists of miR-1273a, miR-1273c, miR-1273d, miR-1273e, miR-1273f, miR-1273g-3p, miR-1273g-5p, miR-1273h-3p and miR-1273h-5p. It was found that the miRNAs of miR-1273 family have binding sites on mRNAs of 33 to 1,074 target genes, with a free energy hybridization of 90% or more of its maximum value.

We found that miR-1273g-3p has 1,330 binding sites on 1,074 mRNAs. Of those, 69 miR-1273g-3p binding sites are located in 5'UTRs, 38 sites are located in CDSs and 1,223 sites are located in 3'UTRs. The mRNAs of seven genes have completely complementary binding sites for miR-1273g-3p ($\Delta G/\Delta G_m = 100\%$).

The mRNAs of 766 genes contain 886 miR-1273f binding sites. Of those, 45 sites are located in 5'UTRs, 40 sites are located in CDSs and 801 sites are located in 3'UTRs. The mRNAs of ten genes have completely complementary binding sites for miR-1273f ($\Delta G/\Delta G_m = 100\%$). The mRNAs of 582 general target genes have pair sites for both miR-1273g-3p and miR-1273f. Of those, 24 mRNAs are located in 5'UTRs, 18 are located in CDSs and 540 are located in 3'UTRs. miR-1273g-3p and miR-1273f binding sites in all of the tested mRNA are located at distance of 12 nucleotides.

We found 449 miR-1273e binding sites on the mRNAs of 413 target genes. Of those, 19 binding sites are located in 5'UTRs, nine sites are located in CDSs and 421 sites are located in 3'UTRs. The nucleotide sequences of paired miR-1273g-3p and miR-1273e binding sites located in the 5'UTR and their adjacent parts have high homology level. The mRNA segments in CDSs containing paired miR-1273g-3p and miR-1273e binding sites also have a high degree of homology. The paired miR-1273g-3p and miR-1273e binding sites are found in the 3'UTR of 300 genes, and they have a high degree of homology. The 3'UTR of target genes have paired miR-1273g-3p and miR-1273a binding sites that are also located in the 5'UTR, with three overlapped nucleotides.

Conclusion: The arranged binding sites of the miR-1273 family are located in the 5'UTR, CDS or 3'UTR of many mRNAs. Five repeating regions containing some of the miR-1273 family's binding sites were found in the 3'UTR of several target genes. The oligonucleotides of miR-1273 family binding sites that are located in CDSs coded for homologous oligopeptides in the proteins of target genes. The nucleotide sequences of mRNAs (99nt) with arranged binding sites for the miR-1273 family were established. Many target genes for miR-1273 family participate in tumorigenesis.

THE FEATURES OF BINDING SITES OF miR-619-5P, miR-5095, miR-5096 AND miR-5585-3P IN THE mRNAs OF HUMAN GENES

Ivashchenko A.T.*, Berillo O.A., Pyrkova A.Y., Niyazova R.E., Atambayeva S.A.
National Nanotechnology Laboratory, *Al-Farabi KazNU, Almaty, Kazakhstan*
e-mail: a_ivashchenko@mail.ru
**Corresponding author*

Key words: miRNA, mRNA, miR-619, binding site, human

Motivation and Aim: The actions of miRNAs on the cell cycle, apoptosis, differentiation, growth and development in animals have been shown. Connections between miRNA expression and the development of various diseases has been established. miRNA concentrations change in cancer and cardiovascular diseases. Metabolic disturbances necessarily change miRNA concentrations in cells. It is possible to normalize some processes using miRNAs. The aforementioned roles do not encompass the full list of the biological processes in which miRNAs participate, which proves the importance of their biological functions. The connections between the majority of miRNAs and their target genes will remain unknown.

Methods and Algorithms: MirTarget program defines the localization of miRNA binding sites in the 5'UTRs, CDSs and 3'UTRs of the mRNAs; it calculates the free energy of hybridization (ΔG, kJ/mole) and the ratio $\Delta G/\Delta G_m$ (%), where ΔG_m equal to the free energy of miRNA binding with completely complementary nucleotide sequence. The binding sites of miRNAs with mRNAs were selected with $\Delta G/\Delta G_m$ ratio of 90% or more.

Results: The binding of 2,563 human miRNAs with the mRNAs of 12,175 human genes was calculated. It was established that miR-619-5p, miR-5095, miR-5096 and miR-5585-3p bind with high affinity to the mRNAs of the 1215, 832, 725 and 655 genes, respectively.

miR-619-5p has 1811 binding sites on 1215 target mRNAs. Of these, 1772 miR-619-5p binding sites are located in 3'UTRs, 26 sites are located in 5'UTRs and 13 sites are located in CDSs. The mRNAs of 197 genes have completely complementary binding sites for miR-619-5p ($\Delta G/\Delta G_m = 100\%$). The mRNAs of 27 genes have four binding sites. Seven genes have five binding sites, and the mRNAs of the *CATAD1, ICA1L, GK5, POLH* and *PRR11* genes have six miR-619-5p binding sites. The mRNAs of the *OPA3* and *CYP20A1* genes have eight and ten binding sites, respectively. All of these sites are located in 3'UTRs. miR-5096 has 997 binding sites on 832 target mRNAs. Of these, 984 miR-5096 binding sites are located in 3'UTRs, nine sites are located in 5'UTRs and four sites are located in CDSs. The mRNAs of 42 genes have completely complementary binding sites for miR-5096 ($\Delta G/\Delta G_m = 100\%$). The mRNAs of the *IP09* gene have four binding sites. The *PRR11* gene have five binding sites. The mRNAs of the *OPA3* and *CYP20A1* genes have six and eleven miR-5096 binding sites, respectively. All of these sites are located in 3'UTRs. We found that 655 target gene mRNAs have 734 binding sites. Fourteen of these binding sites are located in 5'UTRs, eight sites are located in CDSs and 712 sites are located in 3'UTRs. The mRNAs of two genes have completely complementary binding sites for miR-5095. The mRNAs of the *OPA3,* and *SPN* genes each have four binding sites.

The mRNAs of 725 target gene have 844 binding sites for miR-5585-3p. Nine of these binding sites are located in 5'UTRs, two sites are located in CDSs and 833 sites are located in 3'UTRs. The mRNAs of the *CYP20A1* and *GPR155* genes each have four binding sites.

Conclusion: Studied miRNAs have binding sites in the 3'UTRs, CDSs and 5'UTRs. The mRNAs of many genes have multiple miR-619-5p, miR-5095, miR-5096 and miR-5585-3p binding sites. Groups of mRNAs with the ordering of the miR-619-5p, miR-5095, miR-5096 and miR-5585-3p binding sites were established.

MATHEMATICAL MODEL FOR SUBGENOMIC HEPATITIS C VIRUS REPLICATION: IMPACT OF DRUG RESISTANCE EMERGENCE ON LONG-TERM KINETICS OF NS3 PROTEASE INHIBITORS ACTION

Ivanisenko N.[1]*, Mishchenko E.[1], Akberdin I.[1], Demenkov P.[1], Kozlov K.[2], Todorov D.[2, 5], Gursky V.V.[6], Samsonova M.G.[2], Samsonov A.M.[2, 6], Clausznitzer D.[3], Kaderali L.[3], Kolchanov N.A.[1], Ivanisenko V.A.[1, 4]

[1]*Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia;*
[2]*St. Petersburg State Polytechnical University, Petersburg, Russia;*
[3]*Institute for Medical Informatics and Biometry, TU Dresden, Fetscherstraße 74, 01307 Dresden, Germany;*
[4]*PB-soft Llc, Novosibirsk, Russia;*
[5]*Chebyshev lab., St. Petersburg State University, St. Petersburg, Russia;*
[6]*The Ioffe Institute, St. Petersburg, Russia*
e-mail: n.ivanisenko@gmail.com
**Corresponding author*

Motivation and Aim: Hepatitis C virus (HCV) causes hazardous liver diseases, leading frequently to cirrhosis and hepatocellular carcinoma. Direct acting antivirals such as HCV NS3 protease inhibitors significantly improved success rate of HCV therapy. However, the drug resistance of virus remains an important problem. Because of the high error rate of the viral RNA polymerase NS5B, HCV RNA genome is very heterogeneous and rapidly acquires drug resistance.

Methods and Algorithms: We have developed a stochastic model of subgenomic HCV replicon replication, in which the emergence and selection of drug resistant mutant viral RNAs in replicon cells is taken into account. [1] Incorporation into the model of key NS3 protease mutations leading to resistance to BILN-2061 (A156T, D168V, R155Q), VX-950 (A156S, A156T, T54A) and SCH 503034 (A156T, A156S, T54A) inhibitors allows us to describe the long term two-phase dynamics of the viral RNA suppression for various inhibitor concentrations.

Results and Conclusion: We showed that the observable difference between the viral RNA kinetics for different inhibitor concentrations can be explained by differences in the replication rate and inhibitor sensitivity of the mutant RNAs. With the help of our model, we calculated the ratio between the number of mutants in the whole viral RNA pool that are descendants of the preexisting mutants and the number of mutants newly generated in the presence of inhibitor. The results are consistent with the fact that the preexisting mutants of the NS3 protease play the central role in the acquisition of resistance to the NS3 protease inhibitors by the replicon. Interpretation of the results of anti-HCV drug testing on replicon systems will provide opportunity to predict optimal schemes of their usage.

References:

1. Ivanisenko N. *et al*. (2014) PloS one, 9(3), e91502.

MOLECULAR MECHANISMS OF INTERACTION OF TUMOR NECROSIS FACTOR WITH TNF-BINDING ORTOPOXVIRAL PROTEINS CrmB

Ivanisenko N.V.[1]*, Tregubchak T.V.[2], Saik O.V.[1], Ivanisenko V.A.[1], Shchelkunov S.N. [1, 2]
[1]Institute of Cytology and Genetics SB RAS, Novosobirsk, Russia;
[2]State Research Center of Virology and Biotechnology VECTOR, Koltsovo, Novosibirsk region, Russia
e-mail: n.ivanisenko@gmail.com
**Corresponding author*

Motivation and Aim: TNF hyperproduction causes the development of chronic inflammations, including those of autoimmune nature. Recombinant viral proteins inhibiting Tumor Necrosis Factor (TNF) activity could be used for development of new drugs for the inflammatory diseases treatment. Computer simulation of the complexes of interacting proteins and recognition of their interaction patterns allow to plan site directed mutagenesis studies and to predict modified forms of TNF-binding proteins of Orthopoxviruses with improved binding affinity and TNF neutralizing activity.

Methods and Algorithms: Homology models of molecular complexes of human TNF and mouse TNF with TNF-binding domains of cowpox virus and variola virus CRMB proteins were predicted using MODELLER and Rosetta. For analysis of binding free energy 60 ns molecular dynamics simulations (Amber 12) were conducted followed by MM/GBSA protocol. MM/GBSA free energy decomposition protocol was used to identify key residues involved in complexes formation. Affinities of complexes formation were measured experimentally using surface plasmon resonance method (SPR).

Results and Conclusion: It was shown that mouse TNF binds stronger with TNF-Binding Protein CrmB of variola virus or cowpox virus in comparison with human TNF. TNF-Binding Protein CrmB of variola virus had higher affinity to TNF of both organisms than CrmB of cowpox virus. The results of MM/GBSA free energy calculation were confirmed experimentally by surface plasmon resonance measurements. The MM/GBSA free energy decomposition protocol allowed us to reveal the key amino acid residues involved in the formation of the protein complexes, also to explain the observed difference in their formation free energy. Aminoacid substitution ASP63->ASN63 in the sequence of TNF-Binding Domain of CrmB of variola virus was revealed to be energetically favorable. The current work gave a deeper insight into the molecular mechanisms of the interaction of the TNF-binding protein of variola virus and cowpox virus with TNF.

NEW VERSIONS OF THE PDBSITE DATABASE AND PDBSITESCAN TOOL: PREDICTION OF FUNCTIONAL SITES IN THE PROTEIN 3D STRUCTURE

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

Key words: protein functional site prediction, PDBSite database, PDBSiteScan tool

Motivation and Aim: Previously we had developed the PDBSite database (Ivanisenko et al, 2005) containing templates of 3D structures of protein functional sites extracted from the PDB database. Also we created the PDBSiteScan program (Ivanisenko et al, 2004) designed for the prediction of functional sites in the 3D structure of proteins based on the structural alignment of site templates from the PDBSite database with the user-specified spatial structure of the protein. However, only a predefined set of sites could be used for their recognition in protein structure by the PDBSiteScan program. The aim of this work was to create a relational version of the PDBSite database providing a user specified formation of a site templates set to be recognized in the analyzed protein using PDBSiteScan.

Methods and Algorithms: MySQL was used for the development of the relational PDBSite database. The information about 3D structure of protein sites was extracted from the PDB database. The parallel version of the PDBSiteScan program was implemented with the use of MPI.

Results: We developed a new relational version of the PDBSite database containing over 300000 of 3d templates of protein functional sites including binding sites of different ligands. The relational PDBSite database is integrated into the PDBSiteScan program and allows to form a user-specified set of site templates for functional site recognition in spatial structure of the analysed protein.

Work was supported in part by a project VI.61.1.2.

References:

1. V.A. Ivanisenko et al. (2004) PDBSiteScan: a program for searching for active, binding and post-translational modification sites in the 3D structures of proteins, Nucleic Acids Res., **32**, W549-W554.
2. V.A. Ivanisenko et al. (2005) PDBSite: a database of the 3D structure of protein functional sites, Nucleic Acids Res., **33**, D183–D187.

METABOLOME AND TRANSCRIPTOME ANALYSES OF A TOMATO INTROGRESSION LINE CONTAINING A *SOLANUM PENNELLII* CHROMOSOME SEGMENT

Kanayama Y.*, Ikeda H.
School of Agricultural Science, Tohoku University, Sendai, Japan
e-mail: kanayama@bios.tohoku.ac.jp
**Corresponding author*

Key words: metabolome, *Solanum lycopersicum*, *Solanum pennellii*, tomato, transcriptome

Motivation and Aim: Tomato (*Solanum lycopersicum*) is one of the most important fruit crops in the world and its production is flourishing. The tomato fruit is popular as a vegetable and for juice and sauce and contains functional components such as lycopene. There have been several physiological, molecular biological, and genetic studies of tomato as well as conventional horticultural studies, and the whole nucleotide sequence of the tomato genome is now available. Tomato is useful as a fruit crop model because *Arabidopsis thaliana* is not a suitable model for crops with fleshy fruits.

The genetic variation of cultivated tomato species has been reduced by domestication, and tomato wild relatives with useful genes that have been lost from cultivated tomatoes may be useful for breeding. However, most traits associated with fruit quality and environmental stress tolerance are quantitative, and wild relatives have undesirable traits for fruit production, impeding their use in breeding. A set of tomato introgression lines has been developed from a cross between *S. lycopersicum* and wild relative *S. pennellii* for the efficient evaluation and use of useful quantitative trait loci from wild relatives. Each line carries a single *S. pennellii* chromosome fragment in the genetic background of cultivated tomato 'M82', and one line was used as a near-isogenic line for analyzing traits associated with fruit quality and stress tolerance.

Results and Conclusion: One of the introgression lines, IL8-3, carrying a *S. pennellii* chromosome segment on chromosome 8 of *S. lycopersicum* 'M82' was used. IL8-3 fruit showed higher Brix content, probably because of high starch accumulation, than 'M82' fruit. In addition, IL8-3 plants were tolerant to salt stress and blossom-end rot, a major physiological disorder. Furthermore, we performed metabolome and transcriptome analyses to gain a comprehensive understanding of high Brix content and reduced incidence of blossom-end rot in IL8-3, and to find additional useful characteristics. Differences in the levels of many metabolites were observed between IL8-3 and its parental cultivar 'M82' in the middle stage of fruit development as well as in the ripening stage, and different pathways of secondary metabolite synthesis were found between IL8-3 and 'M82'. In contrast, transcriptome analysis showed relatively small differences between IL8-3 and 'M82'. The results indicate that a limited number of genes from wild species may induce a large shift in the metabolome of cultivated tomato fruit.

EVOLUTION OF RUBISCO ENCODING GENES IN PLANTS AND ITS IMPLICATIONS FOR RUBISCO ENGINEERING IN CROPS

Kapralov M.V.*, Whitney S.M.
Plant Science Division, Research School of Biology, The Australian National University, Canberra ACT 0200, Australia
e-mail: maxim.kapralov@anu.edu.au
**Corresponding author*

Key words: adaptive evolution, photosynthesis, plastome transformation, Rubisco

Motivation and Aim: Performance of ribulose-1,5-bisphospate carboxylase/oxygenase (Rubisco) is often one of the limiting steps in photosynthesis, and hence can greatly affect crop yield. In addition to natural Rubisco short-comings, many modern crop cultivars likely possess sub-optimal Rubiscos because they are grown in climates different from the ones experienced by their ancestors. This creates the yield-increasing potential for crops that might be realized by transgenic improvement of their Rubiscos to better fit climates they grow in as well as future climates to come. However, at the present time, chloroplast transformation techniques for most crops are still in their infancy, thus preventing assessment of candidate amino acid replacements that may be responsible for changes in Rubisco properties. Our aim is to better understand evolution of Rubisco encoding genes in plants and to use this knowledge for finding amino acid replacements that might change kinetics of crop Rubiscos in a desired way.

Methods and Algorithms: Maximum Likelihood and Bayesian algorithms implemented in PAML (Yang 2007) were used to analyse signatures of adaptive evolution among Rubisco encoding genes from major lineages of plants to find candidate amino acid replacements that might change Rubisco kinetics. Further, as the first step to assess effects of these substitutions *in planta*, we tested the possibility of assembly of hybrid Rubiscos consisting of large subunits from major cereal, legume, tuberous and vegetable crops and tobacco small subunits in the tobacco transplastomic line, ^{cm}trL, developed by Whitney and Sharwood (2008).

Results: A range of candidate amino acid replacements that could affect Rubisco specificity and velocity of carboxylation was found for the key crops. The hybrid Rubiscos consisting of large subunits from dicot crops and tobacco small subunits assembled within transgenic tobacco into functioning enzymes, while ones consisting of large subunits from monocots did not.

Conclusion: Transgenic tobacco could be used as the Rubisco screening surrogate for dicot crop species to assess the effects of amino acid replacements at sites evolving under positive selection.

References:

1. Z. Yang (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution*, **24:**1586-1591.
2. S.M. Whitney, R.E. Sharwood (2008) Construction of a tobacco master line to improve Rubisco engineering in chloroplasts. *Journal of experimental botany*, **59**:1909-1921.

PLANT MOLECULAR CYTOGENETICS AND NEXT-GENERATION SEQUENCING DATA ANALYSIS

Karlov G.I*., Divashuk M.G., Alexandrov O.S., Razumova O.V., Khuat T.M.L., Kroupin P.Yu.
Centre for Molecular Biotechnology, Russian State Agrarian University – Moscow Timiryazev Agricultural Academy, Moscow, Russia
e-mail: karlovg@gmail.com
**Corresponding author*

Key words: plant molecular cytogenetics, plant genome, DNA repeats, *Cannabis sativa*, *Ricinus communis*, *Triticeae*

Motivation and Aim: With the development of high-throughput sequencing methods a number of draft genomes of plant species has become available. These genomes are presented only by sets of DNA reads and scaffolds. In our study, we consider the possibility of molecular cytogenetics by analyzing draft genomic sequencing data of plant species belonging to families that differ in genome sizes.

Methods and Algorithms: The NCBI available draft genomes of *Ricinus communis*, *Cannabis sativa* and *Triticum* sp. were used. The sequencing reads were clustered on the basis of similarity using RepeatExplorer website (http://repeatexplorer.umbr.cas.cz (as described by Novak et al. 2010), and clusters containing at least 57 reads (representing around 0.01% of the genome) were used for further analysis. To detect tandemly repeated DNA the draft genomes were analyzed by 'Tandem repeat finder' software. The detected repeats were used for fluorescence *in situ* hybridization (FISH) on mitotic metaphase chromosomes.

Results: In this study many high copy number repeats were discovered. FISH analysis revealed subtelomeric, pericentromeric, centromeric, constitutive heterochromatin or chromosome specific tandem repeats distributions. As an example, constitutive heterochromatin specific high copy number 39-bp repeat and chromosome specific tandem 390-bp repeats were discovered in *Ricinus communis*. Studying features of the rcsat39 nucleotide composition and curvature models suggests that it plays a key role in building heterochromatin. The XY sex chromosome determination system of dioecious *Cannabis sativa* was characterized using the obtained genomic data. The novel repeats also allowed us to reveal features of genome organization of some Triticeae species, identify their subgenomic composition and possible evolution. The different groups of retrotransposons were clustered and their genomic organization was studied.

Conclusion: Molecular cytogenetics is a powerful tool for genome mapping and integration of genomic data. By means of combined genomic, bioinformatic and cytogenetic approaches the large scale organization of *Ricinus comnunis*, *Cannabis sativa* and Triticeae genomes were characterized. Our results provide basic molecular cytogenetic data on the structure of genome and could provide a starting point for further studies on genome assembly.

Acknowledgements: This work was supported by Russian Foundation of Basic Research No.13-04-01804 and No.13-04-02116.

References:

1. Novak et al. (2010) Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. *BMC Bioinformatics* 11(1):378–389.
2. Divashuk et al. (2014) Molecular Cytogenetic Characterization of the Dioecious *Cannabis sativa* with an XY Chromosome Sex Determination System. PLoS ONE 9(1): e85118.

CANCER CELL LINE RECOGNITION BY SHOTGUN PROTEOMICS USING CANCER EXOME DATA

Karpova M.A.[1], Karpov D.A.[1], Ivanov M.V.[2], Zgoda V.G.[1], Gorshkov M.V.[2], Moshkovskii S.A.[1]*

[1]*Orekhovich Institute of Biomedical Chemistry RAMS, Moscow, Russia;*
[2]*Institute for Energy Problems of Chemical Physics RAS, Moscow, Russia*
e-mail: smosh@mail.ru
**Corresponding author*

Key words: exome, proteome, shotgun proteomics, cancer cell line, mutant proteins, variant peptides, search engine, MS/MS data

Motivation and Aim: In this work we constructed a sequence database for search of publicly available proteome data of NCI-60 cancer cell lines [1] against recent results of exome sequencing of the same cell lines [2]. The primary objective of this analysis was to test the possibility of cancer cell line recognition based on the sets of variant peptides identified for each of the lines.

Methods and Algorithms: A database of mutant protein variants was generated using NCI-60 cell line exome sequences. Results of bottom-up "shotgun" proteome analysis of the same cell lines were processed using X!Tandem search engine. For in-home validation of the approach, cancer cell lines were provided in blinded manner. Proteomes were subjected to gel separation and proteins were digested by trypsin and analyzed using QExactive analyzer.

Results: A percentage of identified variant peptides corresponding to the own cell line exome was used as a metrics of correct identification. Proteome data contained "deep proteome" and "full proteome" data depending on the fragmentation method (HCD or CID) and the number of fractionation steps. In order to recognize correct cell line, for a particular cell line the identified "correct" variant peptides were matched against all the other lines. The lists of "correct" peptides were always matching best the corresponding exome in all 9 "deep proteome" cell lines. Similar results were obtained for "full proteome" with 50 out of 59 correct matches.

The ability of shotgun proteomics to identify cancer cell lines by variant peptide signatures was further validated using in-home experiments with blinded set of cell lines and algorithms used for previous analysis.

Conclusion: In this study, we have demonstrated that shotgun "deep" proteomics data are sufficient for unambiguous recognition of the cancer cell lines. The task of recognizing unknown cancer tissues is, in some cases, useful for clinical practice.

Availability: All in-home implemented software may be provided upon request.

References:

1. A. Moghaddas Gholami et al. (2013) Global proteome analysis of the NCI-60 cell line panel, *Cell Rep.,* **4**: 609-620.
2. O.D. Abaan et al. (2013) The exomes of the NCI-60 panel: a genomic resource for cancer biology and systems pharmacology. *Cancer Res.,* **73**: 4372-4382.

DIOXIN-MEDIATED REGULATION OF GENES INVOLVED IN CYTOKINES PRODUCTION BY MACROPHAGES

Kashina E.V.[1]*, Oshchepkov D.Y.[1]*, Oshchepkova E.A.[1], Shilov A.G.[1], Antontseva E.V.[1], Furman D.P.[1,2], Mordvinov V.A.[1]
[1]*Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia;*
[2]*Novosibirsk State University, Novosibirsk, Russia*
e-mail: alena_kashina@bionet.nsc.ru; diman@bionet.nsc.ru
**Corresponding author*

Key words: macrophage, dioxin, AhR, cytokines

Motivation and Aim: The environmental pollutant 2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD, dioxin) is the most toxic among the dioxin xenobiotics and induces a broad spectrum of biological responses, including also immunotoxicity and cancer [1]. Macrophages are key regulators of the innate immune response, as well as one of the first types of cells to respond to stress, so the study of the action of dioxin in these cells is important. Dioxin mediates gene expression via AhR/ARNT transcription complex activation, which binds to dioxin responsive elements (DRE) in the regulatory regions of target genes. TCDD acts as a stimulator of some cytokines [2] and our analysis [3] showed that the list of stimulated cytokines is not yet completed. Also the possibility of direct as well as indirect regulation of cytokine synthesis via transcription factors (TFs), through DREs in its promoters was investigated.

Methods and Algorithms: By using the SITECON software package in the regulatory regions of several cytokines (*IL12A*, *IL12B*, *IL4*, *IL10)* and transcription factors (*REL*, *RELA*, *IRF1*, *ATF3*) genes were searched for DREs. EMSA, Real-time PCR and ELISA experiments were performed on HepG2 (human hepatocellular carcinoma cell line), U937 macrophage-like cells and human monocyte derived macrophages to test functionality of predicted sites.

Results: Obtained data demonstrates functional activity of DREs in *IL12B*, *IL4* gene promoters via AhR signal pathway. The mRNA expression dynamics of cytokines genes also evidence the indirect TCDD–mediated modulation of these genes via TFs.

Conclusion: Activation of these genes by AhR ligands can lead to change of regulation key cytokines and this may be a potential pathway for hypersensitivity development and autoimmune diseases induction by dioxin.

Acknowledgements: This work was supported by RFBR (No. 12-04-01736) and budgetary projects VI.60.1.1 and VI.61.1.2

References:

1. Mandal P. (2005) J Comp Physiol B. 175: 221-30.
2. Kerkvliet N. (2009) Biochem Pharmacol. 77(4): 746–60
3. Furman D., et al. (2009) Comput Biol Chem. 33(6):465-8.

COMPARATIVE GENETIC ANALYSIS OF THREE SPECIES OF THE GENUS *TRIBOLODON* (CYPRINIDAE, CYPRINIFORMES) BASED ON SEQUENCE DATA OF MITOCHONDRIAL DNA CO-1 GENE

Katugina A.O.[1, 2]*, Kartavtsev Y.Ph.[1, 2]
[1] *A.V. Zhirmunsky Institute of Marine Biology FEB RAS,Vladivostok, Russia;*
[2] *Far Eastern Federal University, Vladivostok, Russia*
Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
e-mail: anna_katugina@yahoo.com; yuri.kartavtsev48@hotmail.com
**Corresponding author*

Key words: *Tribolodon*, CO-1, mitohondrial DNA

Motivation and Aim: Species of the genus *Tribolodon* belong to subfamily Leuciscinae, family Cyprinidae. Earlier studies, which were based mainly on morphological characters, did not separate southern and northern forms of the Far Eastern dace *T.hakonensis* [1]. The difference thees forms has been found using electrophoretic analysis of protein cording locus Prot-2, Cyt B and D-loop sequences [2,4,5,6]. However, the above-cited authors did not suggest taxonomic status for these genetically different groups of the Far Eastern dace using both morphological characters genetic traits. The aim of the present study is to compare three nominal species of the genus *Tribolodon* using sequence data of mitochondrial DNA (mtDNA) CO-1 gene.

Methods and Algorithms: DNA of three species of the genus *Tribolodon* was extracted. The region of mtDNA that included CO-1 gene was amplified. All sequences were aligned using software package CLUSTAL W. Phylogenetic relationship between species were reconstructed by several techniques using MEGA 6.

Results: According to the CO-1 gene fragment nucleotide sequence signal, southern (Primorsky Kray) and northern (Sakhalin Island) forms of *T.hakonensis* form different clusters with high bootstrap level. Clustering between species of the genus *Tribolodon* is even more explicit and also has high support on gene trees.

Conclusion: Our results well agree with former findings on biochemical genetics and molecular genetics of the Far Eastern dace [4,5,6], in particular, that *T.hakonensis* exhibit very high intraspecific genetic variability. High level of divergence between samples of *T.hakonensis* may suggest that there exist cryptic species in northern and southern areas of the "species" range. These data are preliminary, and require more thorough phylogenetic and population genetic analyses at several molecular markers, both mtDNA and nuclear DNA. Ongoing research will help to assess genetic diversity of *Tribolodon* species, learn their genetic relationships and establish precise taxonomic status for two *Tribolodon hakonenis* forms.

Availability: Dace from waters of the Russian Far East is an object of recreational and commercial fisheries[3]. Expanding our knowledge on population biology and genetic composition of the Far Eastern dace species will be helpful for better understanding life strategies of these anadromous fish, and may also be useful for rational exploitation of fish resources and for conservation of genetically distinct and unique fish stocks.

Acknowledgements: Our thanks to professor Ivankov V.N. (FEFU), Dr. Nikitin V.D. (SakhNIRO), Dr. Turanov S.V. (A.V. Zhirmunsky Institute of Marine Biology FEB RAS).

References:

1. V.V. Sviridov. (2002) Geographic variation of Far Eastern redfins, In: Morphological and genetic divergence and geographic variation of Far Eastern redfins of *Tribolodon* genus. Dissertation council of A.V. Zhirmunsky Institute of Marine Biology FEB RAS. P 115-169.
2. Y.F. Kartavtsev. (2009) Analysis of Sequence Diversity at Two Mitochondrial Genes on Different Taxonomic Levels. Applicability of DNA Based Distance Data in Genetics of Speciation and Phylogenetics // Genetic Diversity. –New York: Nova Science Publishers. P 1-50.
3. N.P. Novikov, A.S. Sokolovsky, T.G. Sokolovskaya, Yu. M. Yakovlev. (2002) Class Osteichthyes- Bony fish, In: The fish of Primorye, Russia, Vladivostok: Far Eastern State Technical Fisheries University Press. P 101-104.
4. H. Sakai (2002) Speciation and dispersal of Tribolodon specias (Pisces, Cyprinidae) around the Sea of Japan // Zoological Science V.19. P 1291-1303.
5. T. Sasaki, Y.P. Kartavtsev, S.N. Chiba, T. Uematsu, V.V. Sviridov, N. Hanzawa (2007) Genetic divergence and phylogenetic independence of Far Eastern species in subfamily Leuciscinae (Pisces: Cyprinidae) inferred from mitochondrial DNA analyses // Genes Genet.Syst. V.82 P.329-340.
6. A.V. Semina, N.E. Polyakova, V.A. Brykov (2006) report RAS V. 407. P. 571-573.

SELF-ORGANIZATION MECHANISMS FOR AUXIN DISTRIBUTION IN THE ROOT APICAL MERISTEM

Kazantsev F.V.[1]*, Chernova V.V.[1], Doroshkov A.V.[1], Omelyanchuk N.A.[1], Mironova V.V.[1, 2], Likhoshvai V.A.[1, 2]
[1]*Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia;*
[2]*Novosibirsk State University, Novosibirsk, Russia*
e-mail: kazfdr@bionet.nsc.ru
**Corresponding author:*

Motivation and Aim: Plant hormone auxin is the main regulator of root apical meristem (RAM) functioning. Auxin maxima coincide with the sites of the RAM initiation and maintenance. Polar auxin transport mediated by PIN family auxin carriers generates nonuniform auxin distributions, which influence cellular dynamics and differentiation. Previously [1] we showed that auxin regulated PIN1 expression with positive and negative feedbacks is a sufficient condition for self-organization of auxin distribution. Here, we show results of our further studies of the outputs from the feedbacks between auxin-regulated transporters expression and auxin distribution pattern.

Methods: We developed a mathematical model, which describes auxin and PIN proteins dynamics in the static 2D cellular root structure. As the root has radial symmetry, we considered only half of the root on the 2D longitudinal section. The modeling area had 20 rows and 4 columns of cells. For the each cell in the model we defined 4 variables (auxin, PIN1, PIN2, PIN3) that involved in the processes of synthesis and degradation of PIN proteins and auxin transport flows between neighboring cells conducted by them. The model was calculated using MGSmodeller and Matlab software. By numerical experiments we have studied two different aspects of auxin distribution self-organization: (1) after the root tip cut and (2) in *pin* mutants. In parallel, we performed the corresponding experiments *in vivo* using DR5-GFP and PIN1::PIN1-GFP *Arabidopsis* plants [2].

Results and conclusions: First, we estimated the model parameters to get the steady state solution that fit well the experimental data under normal conditions. In numerical experiments with the root tip cut, we made calculations from the "cut" cellular layout. As a result, we have shown the key role of shoot-derived auxin flow in the RAM restoration. If the major part (more than half) of the proximal meristem (PM) was cut off, auxin flow from the elongation zone appeared to be too high for redistribution in the remaining part of the PM where cells only accumulated auxin and this did not provide for meristem restoration. In contrast, if the major part of the PM was left in the decapitated root, the auxin transport system was able to cope with the rootward auxin flow, generating an auxin maximum at a distance from the root edge. These numerical results fit very well the experimental data.

We also investigated in the model auxin pattern formation in the root tips of *pin* mutants and found conditions for it. Our results provide an idea how the shoot regulate root development at the level of the RAM self-maintenance.

Acknowledgements: Numerical calculations were performed on supercomputer cluster HKC-30T of Siberian Supercomputer Center SB RAS. The work is partially supported by the Dynasty Foundation grant for young biologists and RFBR grants 11-04-33112-mol-a-ved.

References:

1. Mironova et al., (2010) BMC Systems Biology, 4:98.
2. Mironova et al., (2012) Annals of botany. doi: 10.1093/aob/MCS069.

SPATIALLY DISTRIBUTED BACTERIAL COMMUNITIES: SIMULATION WITH «HAPLOID EVOLUTIONARY CONSTRUCTOR»

Klimenko A.I.*, Matushkin Yu.G., Lashin S.A.
Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia;
Novosibirsk State University, Novosibirsk, Russia
e-mail: klimenko@bionet.nsc.ru
**Corresponding author*

Key words: evolution, simulation, mathematical model, multiscale model, software, bacterial mat, biofilm, prokaryotic community

Motivation and Aim: Most of the prokaryotes live in complex laminated structures – bacterial mats or biofilms. Spatial distribution of species' representatives may play a significant role in local microbial cooperation and competition. Migration processes in populations are important evolutionary factors for such communities. Interacting with the other evolutionary factors, they significantly affect the dynamics of allelic frequencies in populations of the community. Therefore, the impact of spatial distribution of cells and substances on evolution of prokaryotic communities is the subject of theoretical interest.

Methods and Algorithms: We needed a simulation tool, which could describe the object of investigation on the different levels of biological organization, in order to take into account ecological, migration, transport, population, metabolic and genetic processes and their combinations. Therefore, we used for our purpose a software package Haploid Evolutionary Constructor 3D (elaboration of earlier introduced Haploid Evolutionary Constructor [1]).

Results: We analyzed the "Poisoner-prey" models in spatial environments of different structure. It has been shown, that population dynamics and evolution of both "poisoner" and "prey" significantly depend on cells localization, presence/absence of chemotaxis ability and spatial environment structure. We also investigated the model of mutation emergence in spatial distributed inhibitory trophic cycles. The models of this type have expressed correlation between the population dynamics and location of mutation cells origin, as well as the flow direction. The modeling of horizontal gene transfer in fluctuating environmental conditions has shown that system dynamics depends on flow direction, presence/absence of chemotaxis ability and ability to compensate lack of substrates.

Conclusion: It has been shown that spatial structure of bacterial communities and redistribution factors may impact upon population dynamics regimes and evolution of populations in the community.

Availability: http://evol-constructor.bionet.nsc.ru

Acknowledgements: The work is supported by the RFBR grant 13-04-00620, the Stipend of the President of Russian Federation, the Program of the RAS Presidium № 28 "Evolution and biosphere origin".

References:

1. Lashin S.A., Matushkin Yu.G., Suslov V.V., Kolchanov N.A. Computer modeling of genome complexity variation trends in prokaryotic communities under varying habitat conditions // Ecological Modelling, 2012, Vol. 224, No.1, pp. 124-129.

SEEDCOUNTER – MOBILE AND DESKTOP APPLICATION FOR MASS PHENOTYPING SEEDS OF WHEAT

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

Motivation and Aim: The quantitative analysis of wheat seeds morphological features is an important step for breeding high-yielding varieties. In the field the analysis is accomplished manually and it allows just to count the number of seeds in spike and their weight. It appears to be difficult to measure such parameters as seeds length, width, circularity, color. To provide the wheat seeds high-performance phenotyping we implemented a mobile application based on Android (1), that allows to recognize wheat seeds on a sheet of paper and to measure their quantative characteristics. PC version of the application allows to use its functionality to process images obtained from different devices, e.g. scanner.

Methods and Algorithms: Android mobile platform was chosen to develop the application, as it is the widest and open source operation system which is used on most popular mobile devices. To process images obtained from a camera the open library of computer vision Open CV (2) was used. The program input are images of wheat seeds placed on a white sheet of the A4 paper. The program recognizes the paper sheet, wheat seeds on its image and estimates seed morphological characteristics (area and dimensions).

Results: The mobile application obtain images from a built-in camera; the result of analysis can be saved in the device memory or transmitted to a Web-server for storage and further processing. For the PC version program input is the image file obtained by any possible means (scanning, digital photo).

Conclusion: SeedCounter application can be used for wheat seeds high throughput phenotyping in the field conditions when it is difficult or impossible to use other special equipment.

Availability: The SeedCounter application available at Google Play: https://play.google.com/store/apps/details?id=org.wheatdb.seedcounter.

The work is supported by RFBR grant №14-07-31226, budget project VI.61.1.2.

References:

1. Android: Open-source software stack for a wide range of mobile devices and a corresponding open-source project led by Google. http://www.android.com.
2. OpenCV (Open Source Computer Vision Library): Open source computer vision and machine learning software library. http://opencv.org.

TOOLKIT FOR CHIP-Seq BASED COMPARATIVE ANALYSIS OF THE PWM METHODS FOR PREDICTION OF TRANSCRIPTION FACTOR BINDING SITES

Kondrakhin Yu.[1, 2], Valeev T.[1, 2, 3], Sharipov R.[1, 2]*, Yevshin I.[1, 2], Kolpakov F.[1, 2]
[1]*Institute of Systems Biology, Ltd, Novosibirsk, Russia;*
[2]*Design Technological Institute of Digital Techniques, SB RAS, Novosibirsk, Russia;*
[3]*Institute of Informatics Systems SB RAS, Novosibirsk, Russia*
e-mail: shrus79@gmail.com
**Corresponding author*

Key words: transcription factor binding site, ChIP-Seq data, position weight matrix approach, the ROC curve

Motivation and Aim: Despite wide application of the powerful ChIP-Seq technology for experimental identification of transcription factor (TF) binding sites, the computational prediction of the TF-binding sites is also relevant. Many methods for the prediction of the TF-binding sites have been developed over the last decades. Some of them represent position weight matrix (PWM) approach that is the most common and widely used. However, there exists little guidance in the choice among these methods because of a comprehensive comparison of existing methods is still challenging in practice. Thus, the direct use of the ChIP-Seq data for assessing predictive ability of the methods does not seem advisable because of such reasons as the tethered binding [1] or false positive rates of peak detection algorithms.

Methods and Algorithms: We have developed computational toolkit for reliable comparison of prediction methods under condition that unknown fraction of the ChIP-Seq data do not contain genuine TF-binding sites. The proposed toolkit has been designed not only to perform the comparative analysis of site prediction methods but also to compare the peak detection algorithms (such us well-known MACS and SISSRs) for processing the ChIP-Seq data. The toolkit consists of the five independent computational tools implemented with the help of the open source BioUML/ geneXplain plug-in framework (http://www.biouml.org/; http://genexplain.com/).

Results: First, on the base of developed toolkit we have revealed significant dissimilarities between MACS and SISSRs. Second, we have performed comparative analysis of the following three prediction methods that represent PWM approach: common additive model, common multiplicative model and MATCH. The analysis has revealed that MATCH performed significantly worse than two other methods while common additive method outperformed others. It is important to note that this inference is invariant with respect to choice of peak detection algorithm despite the reveled dissimilarities between MACS and SISSRs.

Availability: http://www.biouml.org/; http://genexplain.com/

References:

1. J. Wang *et al*. (2012) Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors, *Genome Res.*, **22:** 1798-1812.

DIRECT AND INVERSE PROBLEMS FOR SYSTEMS WITH SMALL PARAMETER IN KINETICS MODELS

Kononenko L.I.[1, 2]*
[1]*Sobolev Institute of Mathematics SB RAS, Novosibirsk, Russia;*
[2]*Novosibirsk State University, Novosibirsk, Russia*
e-mail: volok@math.nsc.ru
**Corresponding author*

Key words: mathematical modeling, singularly perturbed system, integral manifold, slow surface, inverse problem

Motivation and Aim: Inverse problems are very important for mathematical modeling of biochemical kinetics and for its biological applications.

Methods and Algorithms: The study is based on the method of integral manifolds [1] and its modification to the applied problems [2,3].

Results: A system of ordinary differential equations modeling some catalytic reactions involves slow and fast variables, and it can be regarded as a singularly perturbed system of the form

$$x' = f(x, y, t, \varepsilon), \quad \varepsilon y' = g(x, y, t, \varepsilon), \tag{1}$$

for $x \in R^m$ and $y \in R^n$ with the time derivatives x' and y', an infinitely small positive number ε, and some sufficiently smooth functions f and g. We suppose that $f = \sum_{i+j<p} b_{ij} x^i y^j$ is a polynomial of degree p.

Direct and inverse problems for singular systems with small parameter are stated, which describe catalytic reactions in chemical kinetics. The solution of the direct problem is based on the method of integral manifolds. The inverse problem reduces to finding the coefficients of the polynomial in the right-hand part of the slow equation according to the solution given on the slow surface of the system (1). The above arguments make it possible to obtain existence and uniqueness conditions for the coefficients in the right-hand part of the slow subsystem under the assumption that the condition number of operator f is sufficiently small.

Conclusion: Existence and uniqueness conditions can be obtained for the coefficients of polynomial in the right-hand part of the slow well-conditioned subsystem when solution of the system (1) is given on the slow surface.

Acknowledgements: Supported by SB RAS, Interdisciplinary Project 80, and by RFBR, grant 12-01-00074.

References:

1. Yu.A. Mitropol'skii and O. B. Lykova, (1973) Integral Manifolds in Nonlinear Mechanics. Nauka, Moscow. (Russian).
2. V.M. Gol'dshtein and V. A. Sobolev, (1988) Qualitative Analysis of Singularly Perturbed Systems. Inst. Math., Novosibirsk. (Russian).
3. L.I. Kononenko. (2002) On the smoothness of slow surfaces of singularly perturbed systems. Sib. Journ. of Industrial mathematics. **5 (2):** 109–125. (Russian).

KINETIC SIMULATION OF MITOCHONDRIAL SHUTTLES

Korla K.
School of Life Sciences, University of Hyderabad, India 500046
e-mail: kskorla@gmail.com

Key words: ETC, NADH, Krebs cycle, malate-aspartate shuttle, citrate-pyruvate shuttle

Motivation and Aim: Shuttles are used to carry metabolites across biological membrane to maintain appropriate concentrations on both sides. Besides metabolites, shuttles also facilitate transfer of redox equivalents, across the mitochondrial membrane, as it is impermeable to NADH/NAD^+ couple. The three shuttles reported here work in close association with Krebs cycle to maintain metabolite concentrations and supply redox equivalents to the electron transport chain in mitochondria. In the present work, we aim to make a kinetic model of three shuttles, malate-aspartate, citrate-pyruvate and glycerol phosphate, and integrate it with our earlier model of Krebs cycle and oxidative phosphorylation (Oxphos)[1]. Such a model will enable us to study the system in a kinetic perspective and the effect of several components in the kinetics of the system.

Methods and Algorithms: The set of reactions were simulated using Octave, a high-level, open source interpreted language. It provides the numerical solutions of linear and non-linear problems and performs other numerical experiments[2]. Ordinary differential equation solver was used for simulation calculations and graphical representations (GNUPLOT). Scripts were written individually for the three shuttles and were integrated with the scripts for Krebs cycle and Oxphos.

Results: Simulation curves of shuttles individually and the combined reaction set, shuttles combined with Krebs cycle and Oxphos, were obtained. These curves relate the concentration of various components in the same time frame and can help us to view the impact of a single component on all the other components of the system as a single simulation result.

Conclusion: Kinetic model of shuttles with Krebs cycle and Oxphos have been successfully made. Such a model can be effectively used to study the system and the effects of perturbation in the present system.

Acknowledgements: The author gratefully acknowledges the receipt of Senior Research fellowship from University Grants Commission, India. Thanks are also due to V Lakshmipathi, M Durga Prasad and C K Mitra for their valuable guidance.

References:

1. K. Korla, C.K. Mitra. (2014) Modelling the Krebs cycle and oxidative phosphorylation, Journal of Biomolecular Structure and Dynamics, 32(2), 242-256.
2. GNU Octave: http://www.gnu.org/software/octave/

FUNCTIONAL ANALYSIS OF THE PROMOTER REGION OF THE *Xist* GENE IN MOUSE (*MUS MUSCULUS*)

Korotkova A.M.
Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
e-mail: powerbin@mail.ru

Key words: Xist, X chromosome inactivation, promoter, transcription factors, ER

Motivation and Aim: Gene dosage compensation on sex chromosomes in eutherian mammals is realized by a random inactivation of an X chromosome in females. This process is regulated by the X chromosome inactivation center (Xic) located on the X chromosome and consists of a number of elements. The *Xist* gene is a crucial factor of the inactivation process. It produces a long non-coding RNA that covers all future inactive X chromosome resulting in a heterochromatinisation and transcription silencing of majority of the genes on the inactive X. We are studying the promoter region of *Xist* to investigate the mechanism of its function in the inactivation process.

Methods and Algorithms: Functional analysis of the promoter region of the *Xist* gene was performed by using reporter constructions based on the luciferase gene supported vector under the control of the different promoter regions of the investigated gene and by DNase I *in vitro* footprinting. The promoter DNA sequence was also subjected to a computer analysis in order to find potential binding sites for transcription factors.

Transcription factor ER (estradiol receptor) was predicted in the highly conserved among all mammalian region of the *Xist* promoter. Therefore, we have studied action of estradiol on the *Xist* gene transcription through the site ER. We used reporter constructions containing this site in estradiol-induced cells.

Results and Conclusion: The reporter constructions activity was studied in the different types of female mouse cells: differentiated and undifferentiated; and in the differentiated male cells. During these experiments, we obtained a similar expression pattern in female cells and significant differences between female and male cells. The test results were analyzed and compared with data derived from the computer analysis (we located potential binding sites for transcription factors ER, RAR, SRY etc). It was found that six regions likely to possess either enhancer or silencer activity. The ER (estrogen receptor) was showed in two positions of the promoter region. It was suggested that estrogen might take part in the inactivation process during the embryo development and cells differentiation. To examine this hypothesis the reporter construction activity was tested as well in estradiol-subjected cells. As a result of the experiment activity of one reporter construction was increased almost by a factor of two. Availability of the ER site has been shown by chromatin immunoprecipitation. Therefore, it is possible that estrogen effects the site ER directly or through a mediator.

The ER and other transcription factors may play an important role in the transcription activity of the *Xist* gene and in the inactivation process of the X chromosome.

SEQUENCE-BASED MODEL OF GAP GENE REGULATORY NETWORK

Kozlov K.N.[1]*, Gursky V.V.[2], Kulakovskiy I.V.[3], Muzhichenko V.V.[1], Samsonova M.G.[1]

[1]*St.Petersburg State Polytechnical University, St.Petersburg, Russia;*
[2]*Ioffe Physical-Technical Institute of the RAS, St.Petersburg, Russia;*
[3]*Engelhardt Institute of Molecular Biology, RAS, Moscow, Russia*
e-mail: kozlov_kn@spbstu.ru
**Corresponding author*

Key words: transcription, thermodynamics, reaction-diffusion, drosophila

Motivation and Aim: We applied the systems-level approach to understand the segmentation process in the fruit fly Drosophila embryo. The fruit fly segments are determined simultaneously during the blastoderm stage, just before the onset of gastrulation. The process is controlled by the segmentation gene network.

Methods and Algorithms: We developed a sequence-based model of gap gene regulatory network that controls the segment determination in the early Drosophila embryo [2, 3]. Each of 4 gap genes – *hb*, *Kr*, *gt*, and *kni* – is represented by two state variables: the concentration of mRNA and protein. We analyzed the potential regulatory region for each gene from 12Kbp upstream to 6Kbp downstream of the transcription start site. In DNase accessibility regions we predicted binding sites for 8 transcription factors (TF) – the products of *hb*, *Kr*, *gt*, *kni*, *bcd*, *tll*, *cad*, and *hkb* genes. To model the expression of target gene at the RNA level, we used the thermodynamic approach in the form proposed in [1] and adjusted for the purpose of our study. The target gene expression is proportional to the promoter occupancy that is determined by the concentration of TFs (data extracted from the FlyEx database). Two sets of reaction-diffusion differential equations for mRNA and protein concentrations describe the dynamics of the system [2, 3]. The equations for protein concentrations include the delay parameter to account for the average time between events of transcription initiation and corresponding protein synthesis.

Results: With parameters obtained with the DEEP method [3] the model reproduces the experimental data with high accuracy. We analyzed the impact of selected TF binding sites to the system dynamics and found that in our model the majority of them are actually involved in the correct gap gene expression patterns.

Availability: The software is available from authors upon request.

Acknowledgments: This study was supported by the "5-100-2020" Program of the Ministry of Education and Science of the Russian Federation.

References:

1. X. He, Md A. H. Samee, C. Blatti, and S. Sinha. Thermodynamics-based models of transcriptional regulation by enhancers: the roles of synergistic activation, cooperative binding and short-range repression. PLoS Comput. Biol., 6(9), 2010.
2. J. Jaeger, S. Surkova, M. Blagov, H. Janssens, D. Kosman, K.N. Kozlov, Manu, E. Myasnikova, C.E. Vanario-Alonso, M. Samsonova, D. H. Sharp, and J. Reinitz. Dynamic control of positional information in the early Drosophila embryo. Nature, 430:368–371, 2004.
3. K. Kozlov, S. Surkova, E. Myasnikova, J. Reinitz, and M. Samsonova. Modeling of gap gene expression in Drosophila Kruppel mutants. PLoS Comput. Biol., 8(8): e1002635, 2012.

ELECTROSTATIC PROPERTIES OF BACTERIOPHAGE LAMBDA GENOME AND ITS ELEMENTS: VIRUS VS HOST

Krutinin G.G., Krutinina E.A., Kamzolova S.G., Osypov A.A.*
Laboratory of Cell Genome Functioning, Institute of Cell Biophysics of RAS, Pushchino MR, Russia
e-mail: aosypov@gmail.com
**Corresponding author*

Key words: DNA electrostatic properties, genome physical properties, promoters, promoter up-element, terminators

Motivation and Aim: Bacteriophage lambda (L.) is a classical model object and its genome is extensively studied. However, little is known about physical properties of its genome and its elements. Here for the first time we study their electrostatic potential (EP) properties.

Methods and Algorithms: DEPPDB and its tools [1,2] were used to carry out the analysis.

Results: EP global distribution along L. genome corresponds to the localization of its main regulatory elements in the restricted area with high negative EP. Binding frequency of RNA polymerase to DNA along the genome, measured in direct experiment [3], correlates to the calculated EP. Strong lambda phage promoters have pronounced up-element compared to the absence of it in weak promoters. Promoters with intermediate strength possess weak up-element. Lambda-like phages strong pL promoters possess strong electrostatic up-elements, the sequence texts of which are quite different. As shown earlier [1], strong early T7-like and *E.coli* ribosomal promoters with pronounced up-element have high levels of the EP within it. Strong *E.coli* promoters such as rrnB with eliminated up-element (and thus greatly reduced strength) do not have pronounced EP valleys in the corresponding area. Mutated up-elements with enhanced promoter strength exhibit deep EP valleys and peculiarities of some other physical properties. This may indicate the direct role of EP in promoters functioning. Attachment sites of L. and *E.coli* have high EP for integrase recognition. Rho independent lambda and *E.coli* terminators have the same M-like EP profile, reflecting their palindrome nature, with the same EP scale in three-fold different annotated palindrome length. Rho dependent terminators have no common EP.

Conclusion: Almost all lambda genome elements exhibit E. peculiarities of different kind, that reflect their structural properties and may play role in their biological functioning. Overall genome EP reflects its transcription and host-integration regulation architecture.

Acknowledgements: The authors are grateful to Saveljeva E. G. for technical support and the Institute of Mathematical Problems of Biology of RAS for hosting the Database.

References:

1. A.A. Osypov, G.G. Krutinin, S.G. Kamzolova. (2010) DEPPDB – DNA Electrostatic Potential Properties Database. Electrostatic Properties of Genome DNA, J Bioinform Comput Biol., 8(3): 413-25.
2. A.A. Osypov, G.G. Krutinin, S.G. Kamzolova. (2012) DEPPDB – DNA Electrostatic Potential Properties Database. Electrostatic Properties of Genome DNA elements, J Bioinform Comput Biol, 10(2) 1241004.
3. Y. Harada et al. (1999) Single-Molecule Imaging of RNA Polymerase-DNA Interactions in Real Time, Biophys. J., 76(2): 709–715.

TRANSCRIPTION FACTORS AND ELECTROSTATIC AND OTHER PHYSICAL PROPERTIES OF THEIR BINDING SITES

Krutinina E.A., Krutinin G.G., Kamzolova S.G., Osypov A.A.*
Laboratory of Cell Genome Functioning, Institute of Cell Biophysics of RAS, Pushchino MR, Russia
e-mail: aosypov@gmail.com
**Corresponding author*

Key words: TF, transcription factors binding sites, TF electrostatics, DNA electrostatic properties, genome physical properties, promoters

Motivation and Aim: Transcription regulation is well known to be influenced by physical properties of genome DNA, especially by its electrostatic interactions with different regulative proteins. We studied different families of transcription factors in different prokaryotic taxa together with their binding sites to reveal the role of these properties in the transcription regulation proteins binding.

Methods and Algorithms: DEPPDB and its tools [1, 2] were used to carry out the analysis.

Results: Electrostatics (E) distribution on TF surface reflects the DNA E of their binding sites. The averaged profiles of the DNA E potential aligned around the binding sites centers exhibit the pronounced rise in the negative potential value with the characteristic profile in the consensus area (often being a palindrome). The extensive (around 100-300 bp long), symmetrical overall potential rise can not be explained by the influence of the consensus alone and reflects the sequence organization of the flanking regions, contributing to the high potential area formation. Apparently this sequence organization was selected evolutionary to support the binding site recognition by the regulation protein molecule and its retention. TF have strong electropositive patch at the DNA binding area, that is surrounded by mostly electronegative or neutral surface. This may lead to proper orientation of TF molecule and facilitate DNA TF site recognition and binding.

The high potential area is relatively AT-enriched, which is reflected in that different other physical properties, especially energy-related, exhibit similar behavior, though the size and parameters of peculiarities are different. This may facilitate binding and accompanying DNA bending. The same overall properties, though vary in particular details, are typical to binding sites of other families of transcription factors in a diverged range of bacterial taxa.

Conclusion: The role of electrostatics in the regulations of cell genome functioning and evolution is significant and universal. This may facilitate the horizontal gene transfer and adaptation of new regulatory circuits for the pan-genome evolution.

Acknowledgements: The authors are grateful to Saveljeva E.G. for technical support and the Institute of Mathematical Problems of Biology of RAS for hosting the Database.

References:

1. A.A. Osypov, G.G. Krutinin, S.G. Kamzolova. (2010) DEPPDB – DNA Electrostatic Potential Properties Database. Electrostatic Properties of Genome DNA, J Bioinform Comput Biol., 8(3): 413-25.
2. A.A. Osypov, G.G. Krutinin, S.G. Kamzolova. (2012) DEPPDB – DNA Electrostatic Potential Properties Database. Electrostatic Properties of Genome DNA elements, J Bioinform Comput Biol, 10(2) 1241004.

COMPUTER ANALYSIS OF CHROMOSOME CONTACTS REVEALED BY SEQUENCING

Kulakova E.V.[1*], Bryzgalov L.O.[1], Orlov Y.L.[1, 2, 3], Li G.[3], Ruan Y.[3]
[1]*Novosibirsk State University, Novosibirsk, Russia;*
[2]*Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia;*
[3]*Genome Institute of Singapore, Singapore*
e-mail: kylakovaekaterina@gmail.com
**Corresponding author*

Key words: sequencing technologies, chromosome contacts, ChIA-PET, transcription

Motivation and Aim: Studying of higher-order chromosomal organization for transcription regulation in eukaryotes is challenging problem. Evidence from in situ fluorescence studies in the last decade suggests that transcription is not evenly distributed and is concentrated within large discrete foci in mammalian nuclei, raising the possibility that genes are organized into "transcription factories" containing RNA polymerase II and other protein components. Chromosome Conformation Capture (3C) - method of analyzing three-dimensional structure of chromosomes in the cell nucleus. The method can not identify interacting protein. ChIA-PET method to determine the spatial structure of chromatin associated with a specific protein. Does not require knowledge of the DNA sequence in the interaction, completely depends on the specificity of antibodies.

Methods and Algorithms: Using genome-wide Chromatin Interaction Analysis with Paired-End-Tag sequencing, we mapped long-range chromatin interactions associated with RNA polymerase II in human cells and uncovered widespread promoter-centered interactions [1]. These interactions could be further aggregated into higher-order clusters, in which proximal and distant genes are engaged through promoter-promoter interactions. We have compared gene location and chromosome interacting sites. Same approach was used for ER-mediated interactome [2]. In addition, we have developed a computer program statistical data of genes and chromatin domains, analysis of experimental data ChIP-seq, Hi-C. Using genomic annotations the UCSC Genome Browser.

Results: We show that most genes with promoter-promoter interactions are highly active and could transcribe cooperatively, and that some interacting promoters could influence each other, implying combinatorial complexity of transcriptional controls. We found enrichment of ChIP-seq defined transcription factor binding sites from ENCODE project in human genome in spatial proximity to chromatin bound contacting sites. As well as were identified gene ontology categories of genes lying on the domain boundaries.

Conclusion: The study provides insights into the three-dimensional basis of gene transcription activity and new approaches for transcription regulation modeling.

References:

1. G. Li et al. (2012) Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation, *Cell*, **148**(1-2):84-98.
2. M.J. Fullwood et al. (2009) An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature*, **462**(7269): 58-64.

COMPUTER DATA ANALYSIS OF GENOME SEQUENCING BY TECHNOLOGY ChIP-seq AND Hi-C

Kulakova E.V.[1], Podkolodnaya O.A.[2], Serov O.L.[2], Orlov Y.L.[1, 2]*
[1]*Novosibirsk State University, Novosibirsk, Russia;*
[2]*Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia*
e-mail: orlov@bionet.nsc.ru
**Corresponding author*

Key words: ChIP-seq, ChIA-PET, Hi-C, chromosome contacts

Motivation and Aim: Currently there is an active development of high-tech DNA sequencing, including using Chromatin Immuno-Precipitation (ChIP-seq), which leads both to the rapid accumulation of a large number of genome-wide information and qualitatively new tasks of bioinformatics, requiring the development of new computer tools. The aim of this work was to develop a statistical data of the computer program of genes and chromatin domains, analysis of experimental data ChIP-seq, Hi-C in the mouse genome (stem cells) and in the human genome [1], [2].

Methods and Algorithms: We use data on the location spatial domains in the human genome and mouse genome[3] and unpublished data Hi-C that is received at the Institute of Cytology and Genetics SB RAS on the methodology presented in the Battulin's work [4]. As standard genomic annotations we used the UCSC Genome Browser - integrated environment for displaying any requested part of the genome at any size, along with dozens of standardized annotations (http://genome.ucsc.edu).

Results: Identified genes that lies on the spatial boundaries of chromosomal domains. Gene Ontology categories were identified with the help of Internet resources of DAVID (http://david.abcc.ncifcrf.gov/). Coexpression communications were analyzed using internet resource String-db (http://string-db.org/).

Conclusion: Genes on domain boundaries represent approximately 2.3% of the total number of genes. During the analysis of gene ontology categories was observed that the most of the genes that lie on the domain boundaries are responsible for allocating the phosphoprotein. The most important gene ontology categories based on the significance factor that is connected with the plasmatic membrane. Which corresponded to 1/4 of these genes.

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. M.J. Fullwood et al. (2009) An oestrogen-receptor-alpha-bound human chromatin interactome, *Nature*. V. 462. – N. 7269. – P. 58-64.
3. J.R. Dixon et al. (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions, *Nature,* V. 485. – N. 7398. – P. 376-380.
4. N.R. Battulin, et al. (2012) 3C-based methods for 3D genome organization analysis*, Vavilov journal of genetics and breeding*, V. 16. – № 4/2. – C. 872-878.

CARDIAC MECHANICS, CALCIUM OVERLOAD AND ARRHYTHMOGENESIS

Kursanov A.[1]*, Solovyova O.[1, 2], Katsnelson L.[1], Medvedev K.[2], Vasilyeva A.[1, 2], Vikulova N.[1], Markhasin V.S.[1, 2]
[1]*Institute of Immunology and Physiology UB of RAS, Yekaterinburg, Russia;*
[2]*Ural Federal University, Yekaterinburg, Russia*
e-mail:alexander.kursanov@gmail.com
**Corresponding author*

Key words: calcium overload, rhythm disturbances, cardiac mechanics

Motivation and Aim: It is well-known that Ca^{2+} overload may cause cardiac arrhythmia. However, possible contribution of the mechanical factors to the arrhythmia development in Ca^{2+}-overloaded cardiomyocytes has been insufficiently addressed. Earlier we have developed a mathematical model of cardiomyocyte electro-mechanical function [1] that predicted a significant role of the intra- and extracellular mechanical factors in arrhythmogenesys. Model prediction was verified in experiments on papillary muscles from the right ventricle of guinea pigs overloaded with calcium [1].

Methods and Algorithms: We utilized the cellular model to study effects of the electromechanical coupling between cardiomyocytes in a 1D heterogeneous muscle strand formed of 90% of normal (N) cardiomyocytes and 10% of sub-critical (SC) cardiomyocytes with decreased Na^{+}-K^{+} pump activity. Single SC-cardiomyocytes did not demonstrate spontaneous activity during isometric contractions at a reference length. Regular fiber twitches at the reference initial cell length were induced by 1 bps electrical stimulation applied at an edge of the strand. Excitation spread along the tissue via electro-diffusional cell coupling followed by cell contractions and force development in the fiber.

Results: Mechanical interactions between N- and SC-cells in the tissue resulted in the spontaneous activity emerged in the SC-zone between the regular stimuli. If the excitation wave spread from SC- to N-region, the SC-cells developed delayed after-depolarizations (DAD) that caused a slowly developing beat-to-beat decrease in the force of fiber contraction. If the excitation spread in opposite direction, DAD in the SC-cells induced reflected downward excitation waves capturing the normal region and followed by extrasystoles in the whole fiber.

Conclusion: The results obtained in the model suggest that ectopic activity may emerge in a sub-critical myocardial region, e.g. comprising cardiomyocytes with moderately depressed N+-K+ pump, due to its mechanical interactions in the myocardial tissue. Moreover, such ectopic zone may expand by capturing normal regions in myocardium via the electro-mechanical coupling between cardiomyocytes.

Acknowledgements: Supported by the RFBR (14-01-00885, 14-01-31134), by Presidium of UB RAS (12-M-14-2009, 12-П-4-1067) and by UrFU (Act 211 of RF Government #02. A03.21.0006).

References:

1. Katsnelson L.B. et al. (2011) Contribution of mechanical factors to arrhythmogenesis in calcium overloaded cardiomyocytes: Model predictions and experiments, Progress in Biophysics and Molecular Biology, V.107(1), 81-89.

PROTEOMICS EXTREME IMPACTS AS A TOOL FOR SYSTEMS BIOLOGY

Larina I.M.[1]*, Ivanisenko V.A.[2], Nikolaev E.N.[3]
[1]*Institute for Biomedical Problems of RAS, Moscow, Russia;*
[2]*Institute Citology and Genetics SB RAS, Novosibirsk*, Russia;
[3]*Emanuel Institute for Biochemical Physics, Moscow*, Russia
e-mail: irina.larina@gmail.com
**Corresponding author*

Key words: space flight, proteomics, bioinformatics

Systems biology (SB), as a scientific discipline that studies the biological systems from the standpoint of holism involves gathering dynamic global data sets, together with phenotypic data from different levels of the hierarchy of biological information to identify and explain the mechanisms of emergent properties of the system. Environmental gravitational physiology and SB are natural symbionts. SB is based on: (1) the ability to measure all the variables of interest (OMICs), (2) the presence of the conceptual framework for the interpretation of data (models available) and (3) applying the methods of disturbances in the experiment. Such effects on the body, which can disturb homeostasis, provide an opportunity to recognize the mechanisms for maintaining the stability of the internal environment, the adaptive potential of the organism. When the object of study is a healthy person, a list of methods (conditions), ethically permitted and available for experimental exposure is relatively short. Environmental studies, including conditions extreme temperatures and hyperbaric hypoxia, as well as space flight - provide the unique platform for studying human physiology from systemic perspective. Space flight is unprecedented experience of healthy human adaptation to extreme conditions in the history of the Earth evolution. Phenomenology of changes apparent in the major physiological systems of the body, induced by space flight conditions, is well documented at the physiological level. However, the molecular mechanisms for the formation of new states of physiological systems are practically unknown. The task of evaluating the interaction of molecular components of the biological system, and the integration of this information into the network or pathway, which can be used to create predictive models of system behavior - is a serious challenge for researchers, who are developing bioinformatics methods in proteomics analysis of extreme conditions.

DEC: SOFTWARE TOOLS FOR SIMULATION EVOLUTION IN DIPLOID POPULATIONS

Lashin S.A.*, Matushkin Yu.G.
Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia;
Novosibirsk State University, Novosibirsk, Russia
e-mail: lashin@bionet.nsc.ru
**Corresponding author*

Key words: evolution, simulation, mathematical model, multiscale model, software

Motivation and Aim: Evolutionary processes are involved in all levels of biological organization: from molecular-genetic systems to populations and communities. Mathematical modeling is one of the major tools for theoretical studies of evolutionary processes. We present a software package "Diploid evolutionary constructor" (DEC), which provides tools for simulating population genetic and ecological processes in populations of complex diploid organisms.

Methods and Algorithms: The DEC is based on hierarchical approach of construction of multiscale computational models, previously developed by us for simulation of evolution of prokaryotes [1]. In the DEC, we also use the agent-based modeling approach, sequence-modeling techniques, numerical calculations of ODE, and discrete-automaton models (the last two are used for the simulation of gene networks evolution). The DEC is developed in C++ with the use of MPI, OpenMP, Cuda (for high-performance versions of computational core), and QT (for graphical user interface). The support of plugins in DEC is provided and documented.

Results: The DEC may be used in several cases: (1) agent-based modeling of classical population genetic tasks; (2) forward-time simulation of sequences' evolution; (3) simulation of evolution of gene networks represented as ODE systems or discrete-automaton models. Basing on estimations of population genetic parameters obtained in [1], a series of models of evolution of modern human ancestors inside the territory of Eurasia was implemented in the DEC. Using the OpenMP version of the DEC, we obtained the acceleration of calculations about 4-4.5-fold (desktop with 6-core AMD Phenom processor). Several possible evolutionary scenarios have been analyzed.

Conclusion: The DEC is the tool for simulation and analysis for a wide range of evolutionary problems, including simulation of evolutionary-genetic processes in ancient human populations.

Availability: http://evol-constructor.bionet.nsc.ru

Acknowledgements: The work is supported by the RFBR grant 12-07-00671-a, the Stipend of the President of Russian Federation, the Program of the RAS Presidium №28 "Evolution and biosphere origin".

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.

FORWARD-TIME SIMULATION OF EVOLUTIONARY PROCESSES IN ANCIENT POPULATIONS USING THE DIPLOID EVOLUTIONARY CONSTRUCTOR

Lashin S.A.*, Suslov V.V., Gunbin K.V.
Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia;
Novosibirsk State University, Novosibirsk, Russia
e-mail: lashin@bionet.nsc.ru
**Corresponding author*

Key words: human evolution, Homo denisova, simulations of evolution, ancient populations

Motivation and Aim: Reconstruction of admixture history in the course of modern human evolution is the challenge of human genetics. Recent discoveries of *Homo denisova (H.d.)* and *Homo neanderthalensis (H.n.)* genetic investment into modern human genome led to significant revision of views on the history of *Homo sapiens (H.s.)* origin. Using the methods of mathematical modeling and computer simulations, we analyze various scenarios of admixture between *H.n., H.d.* and *H.s.*, based on mitochondrial markers, which allow us to split the investment of maternal and paternal populations.

Methods and Algorithms: The models were constructed and analyzed with the use of the "Diploid evolutionary constructor" software package.

Results and Conclusion: Models contain two populations – "modern humans" and "ancient humans: *H.d.*/*H.n.*». Each individual possesses "mitochondrial genome" which is represented as a set of loci of specified length. Starting sequences are generated stochastically. Starting populations are homogeneous. The ancients differ moderns only at certain part of loci. In the course of evolution, neutral mutations occur with a probability of 10^{-6} per nucleotide/generation. In our model, we consider a recombination between maternal and paternal mtDNA with a certain probability (default value is 1/10000), origin of nuclear mitochondrial DNA's (numt's) and mutual migrations between populations. The program generates output files with "mitochondrial genome" sequences for all individuals of both populations. The number and length of numt's is estimated. Let the migrating population of *H.n.* consists primarily of male individuals, and reaches the areal of nonmigratory *H.s.* population with equal sex ratio. The longer is the time of contact between populations, the higher is the exchange of genetic material between populations (origin of paternal numt's and hybrid mitochondrial DNA). After finishing the contact, each population evolves independently, affected by population dynamics factors, which in turn affects admixture tracks in nonmigratory *H.s.* population. Consequently, the study the above-described computer model allows us both to prioritize admixture markers by their in time, and to differentiate mixing populations on paternal (ecdemic) and maternal (endemic).

Acknowledgements: The work is supported by the RFBR grant 12-07-00671-a, the Stipend of the President of Russian Federation, the Program of the RAS Presidium № 28 "Evolution and biosphere origin".

HAPLOID EVOLUTIONARY CONSTRUCTOR 3D: A TOOL FOR SIMULATION OF SPATIALLY DISTRIBUTED PROKARYOTIC COMMUNITIES

Lashin S.A.*, Klimenko A.I., Mustafin Z.S., Chekantsev A.D., Zudin R.K., Matushkin Yu.G.
Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia;
Novosibirsk State University, Novosibirsk, Russia
e-mail: lashin@bionet.nsc.ru
**Corresponding author*

Key words: evolution, simulation, mathematical model, multiscale model, software, bacterial mat, biofilm, prokaryotic community

Motivation and Aim: Existence and living of unicellular haploid organisms, bacteria and archaea, is laid, as a rule, in complex spatially distributed systems like biofilms, bacterial mats etc. Lots of prokaryotes in them, form multi-specie communities with organized trophic structure, and most of them can not be cultured in artificial conditions, beyond their natural habitats. That is why tools for mathematical and computer modeling prove to be useful in studying the peculiarities in functioning and evolution of prokaryotic communities. We present a novel cross-platform software tool Haploid Evolutionar Constructor 3D (HEC-3D) designed for simulation of prokaryotic communities, their functioning and evolution. HEC-3D is the further development of the HEC software [1].

Methods and Algorithms: We use hybrid agent-based modeling, where the base agent is the polymorphic population of prokaryotic cells. High-level processes (cells reproduction, transport processes, mutations, gene horizontal transfer and loss of genes) are described with an event-driven model, while low-level processes (for example, cellular metabolism) are described with the standard numerical ways (like ODE integration). Computational core is written in pure C++ while the GUI uses QT. We have also implemented several multithread versions of the computational core with the use of OpenMP, Cuda and MPI.

Results: HEC-3D provides simulation and analysis of functioning and evolution of prokaryotic communities taking into account complex spatial organization and nonhomogeneity of their habitats, which are conditioned by such processes as diffusion and chemotaxis. Metabolism and reproduction of cells is determined by associated gene networks. We interpret the change of parameters in these networks as mutations, and structural changes as horizontal gene transfer or genetic loss. Phage infections can be simulated in the HEC-3D in both lysogenic and lytic pathways.

Conclusion: HEC-3D presents system wide studying of complex spatially distributed prokaryotic communities via modeling of such levels of biological organization as genetic, metabolic, population and ecological.

Availability: http://evol-constructor.bionet.nsc.ru

Acknowledgements: The work is supported by the RFBR grant 13-04-00620, the Stipend of the President of Russian Federation, the Program of the RAS Presidium № 28.

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.

MODELING OF CELL DYNAMICS IN THE ROOT APICAL MERISTEM WITH DYNAMICAL GRAMMAR

Lavrekha V.V.[1*], Omelyanchuk N.A.[1], Mironova V.V.[1, 2]
[1]*Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia;*
[2]*Novosibirsk State University, Novosibirsk, Russia*
e-mail: vvl@bionet.nsc.ru
**Corresponding author*

Key words: RAM, cell dynamics, Dynamical Grammar, auxin, cytokinin

Background: Root apical meristem (RAM) is the plant stem cell niche which provides for the formation and continuous development of the root. In the root tip along the central axis there are several zones with principally different cell dynamics: columella, quiescent center, proliferation domain, transition domain and elongation zone. It is known that in the root tip, auxin and cytokinin distributions are nonuniform with maxima and gradients. The distal maxima of auxin and cytokinin in the root tip maintains the stem cell niche. Here we propose that patterning of the plant hormones can be sufficient condition for the zones emergence and maintenance. This hypothesis was tested by mathematical modeling with Dynamical grammar [1].

Material and Methods: Previously [2] we developed a model to study auxin patterning in the stem cell niche. In the model cell divisions were regulated by auxin and hypothetical morphogen[2]. The model was created using package Plenum [3] in Mathematica. The cell dynamics was modeled by Dynamical Grammar rules that specify a weak (breakable) spring potential function between neighboring cells[1]. Here we extended the model with the rules for cytokinin effects on cell dynamics. We substituted the *Division Factor* by cytokinin and auxin and considered their effects on cell cycle.

Results and Conclusions: In the resulting model we observed simultaneous formation of (1) auxin distribution; (2) cytokinin distribution; (3) zonation of the root along the central axis with respect to cell growth/division. We found the model parameters under which the proliferation domain was stable in development. We assume that this model will allow us to study an effect of auxin and cytokinin on cell dynamics in the plant.

References:

1. E. Mjolsness. (2010) Towards Measurable Types for Dynamical Process Modeling Languages.Electr. Notes Theor. Comput. Sci. 265: 123-144.
2. V.V. Mironova et al. (2010) A plausible mechanism for auxin patterning along the developing root. BMC Systems Biology, 4:98.
3. http://computableplant.ics.uci.edu/theses/guy/Plenum.html

SEARCH FOR EVOLUTIONAL INVARIANTS BY RANK DISTRIBUTION OF GENE DENSITY IN HOMINIDS

Lenskiy S.V.*, Lenskaya T.I.
Ural Federal University, Ekaterinburg, Russia
e-mail: sergeylensky@rambler.ru
**Corresponding author*

Key words: evolutional invariants, gene density, rank distribution, power law, hominid genomes

Motivation and Aim: Rank distribution is useful for investigating nucleotide sequences [1, 2]. Application of rank distribution to complementary pairs gave an opportunity to find a pattern for assessment of order in eukaryotic genomes of different origin, to construct theoretical model of the most statistically ordered genome and to identify the real representative with these properties – Arabidopsis thaliana [3]. In this research rank distribution was used for studying the density of genes in the genomes of hominids.

Methods and Algorithms: The genome annotation reports for the hominid genome projects available now on NCBI portal {http://www.ncbi.nlm.nih.gov/genome/} are suitable for estimating the density of genes on autosomes: Homo sapiens GRCh38, Pan_troglodytes-2.1.4, Gorilla gorilla RefSeq Genome, Pongo abelii P_pygmaeus_2.0.2. The gene density of an autosome is the ratio of a number of genes located on the autosome per size of the autosome in million base pairs (Mb). The gene density values of autosomes in a genome were arranged in rank order. We used the least squares method to find an approximation curve for the rank distribution. To compare the observed values of gene density to the modeled values we used the coefficients of determination (R2).

Results: The study of the rank distribution of gene density on autosomes in the hominid genomes showed this distribution could be approximated by a power function: $f(x) = a*x^b$, where a and b were parameters of an approximation curve for a genome. Approximation of genome data from different sources with high values of coefficient of determination indicated a presence of a close connection between the power law and the data.

Conclusion: The pattern of gene density distribution is probably related to an allocation of functional load to autosomes in the hominid genomes. The differentiation of chromosome "employment" may differ by several times within the genome. The greatest burden in the genomes of hominids is experienced by chromosome 19. It is the absolute leader in the density of genes for the observed genomes. Perhaps it implies the special role of this chromosome. Chromosome 20 and chromosome 9 keep constant positions in the rank distributions of hominids. It is likely to be evolutional invariants for the whole family. The second found invariant is parameter b of the power law. The applicability of the power law for a derivative (gene density) is an opportunity for comparison and unification of genome researches. Rank research tools can be and ought to be constantly enhanced.

References:

1. R.N. Mantegna et al. (1994) Linguistic Features of Non-coding DNA Sequences, *Phys. Rev. Lett.,* **73**: 3169-3172.
2. P.A. Tsonis, A.A. Tsonis. (2002) Linguistic features in eukaryotic genomes, *Complexity*, **7(4):** 13-15.
3. S.V. Lenskiy, T.I. Lesnova. (2007) Searching of Statistically Ordered Eukaryote Genome, *ISMB/ECCB 2007*, poster session (O18), 21–25 July, Vienna, Austria.

THE PIPELINE FOR COMPOSITE REGULATORY ELEMENTS PREDICTION

Levitsky V.G.*, Mironova V.V.
Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
e-mail: levitsky@bionet.nsc.ru
**Corresponding author*

Key words: transcription factor binding sites recognition, motif discovery

Motivation and Aim: High-throughput sequencing technologies applied to whole genomes, transcriptomes and transcription factors binding sites (TFBS) require development of computer tools for their integrative analysis. To study TFBS one needs to combine the information about distribution of potential TFBS with gene expression data and other functional genomics data. Here we present a pipeline integrating this information applied to study composite auxin responsive elements (AuxRE). The hormone auxin is a major regulator of plant growth and development, and AuxRE prediction can provide the substantial support for functional annotation of a gene.

Methods and Algorithms: We compiled the sample of 25 experimentally validated AuxRE. For AuxRE prediction we applied our tools optimized Position Weight Matrix (oPWM) and SiteGA since they represented very different models of TFBS [1]. A genetic algorithm based approach MotiGA [2] is employed to detect overrepresented motifs in the nearest flanking sequences of AuxRE (total length 106 nt). Hence, we suppose that there are several types of composite elements consisting of anchor TFBS with a certain coupling motif, which relate to experimentally detected gene expression pattern. To test this hypothesis we processed a number of microarray datasets on the changes of gene expression as response to auxin action for *Arabidopsis thaliana* genes. These datasets were used to test the hypothesis on statistical significance of dependence between the prediction of the composite element in a gene regulatory region and the gene response to auxin detected in an experiment; we estimated this significance by Welch's t-test for arcsine square-root transformed proportions.

Results and conclusions: oPWM and SiteGA methods were trained on the experimentally validated AuxREs. The most conservative AuxRE hexamers (best matches to TGTCTC) were masked with N_6 in AuxRE alignment and next were applied to MotiGA for motif discovery. We revealed three types of overrepresented motifs. Potential simple and composite AuxREs were predicted in upstream regions of *A. thaliana* genes. Estimation of significance for dependences between the presence of a specific composite element with the change of gene expression approved that for all three types of composite AuxREs were significantly associated with auxin response. Moreover, this analysis allowed to reveal the link between specific spacer and/or orientation of the motifs in the composite AuxREs with auxin response pattern. The pipeline is a powerful tool to screen for composite TFBS associated with the specific expression pattern.

Availability: http://wwwmgs.bionet.nsc.ru/mgs/programs/motiga/

References:

1. V.G. Levitsky et al. (2007) Effective transcription factor binding site prediction using a combination of optimization, a genetic algorithm and discriminant analysis to capture distant interactions, *BMC Bioinformatics*, **8**: 481.
2. V.G. Levitsky. DNA motif search by genetic algorithm. Proceedings of the eighth international conference on bioinformatics of genome regulation and structure \ System biology (BGRS\SB'2012), 2012, 182.

LINEAGE-SPECIFIC PROCESSES OF GENOME DIVERSIFICATION

Liberles D.A.
Department of Molecular Biology, University of Wyoming, Laramie, WY, USA;
From October 1, 2014: Department of Biology, Temple University, Philadelphia, PA, USA
e-mail: liberles@uwyo.edu

Keywords: mechanistic models, phylogenetics, molecular evolution, gene duplication, evolutionary systems biology, metabolic pathways

Motivation and Aim: Standard bioinformatics pipelines for comparative genomic data analysis typically include steps like multiple sequence alignment, phylogenetic tree reconstruction, calculation of the ratio of rates of nonsynonymous to synonymous nucleotide substitution (dN/dS), and gene tree-species tree reconciliation. Two projects are presented that describe an increasing level of mechanistic (evolutionary and biochemical) thought that can ultimately be applied in such pipelines.

Methods, Results, and Conclusion: Models for duplicate gene retention, if a biological interpretation is given, assume that the process is evolutionarily neutral and do not consider functional or mechanistic underpinnings. A model for duplicate gene retention that would allow for mechanistic differentiation between nonfunctionalization, neofunctionalization, subfunctionalization, and dosage balance based upon differences in expected time-dependent retention profiles is presented.

A second part of the talk focuses on the evolution of metabolic pathways. Using simulations, the co-evolution of enzyme concentration, Kcat and Km (and specificity) in metabolic pathways is evaluated over several network topologies, a range of mutation rates (including duplication and loss rates) and selective regimes using a kinetic model. The model pathway initially used is a simplified version of the methylglyoxal pathway. Biochemical thought is that there is strong selective pressure to preserve rate limiting steps at early steps in pathways to prevent wasteful protein expression and to prevent the buildup of potentially deleterious intermediates in large concentrations. Evolutionary thought is that mutation-selection balance is an important aspect of the evolutionary dynamics of pathways, leading to evolutionary instability of rate limiting steps. These dynamics are examined under selective regimes that also include either negative or positive selective pressures on pathway flows and outputs.

Together, these approaches reflect an in-progress strategy for increasing biochemical and evolutionary realism in bioinformatics.

TAT-REV REGULATION OF HIV-1 REPLICATION: MATHEMATICAL MODEL PREDICTS THE EXISTENCE OF OSCILLATORY DYNAMICS

Likhoshvai V.A.[1, 2], Khlebodarova T.M.[1], Bazhan S.I.[3]*, Gainova I.A.[4], Chereshnev V.A.[5], Bocharov G.A.[6]

[1]*Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia;*
[2]*Novosibirsk State University, Novosibirsk, Russia;*
[3]*State Research Center of Virology and Biotechnology «Vector», Koltsovo, Russia;*
[4]*Sobolev Institute of Mathematics SB RAS, Novosibirsk, Russia;*
[5]*Institute of Immunology and Physiology, Ural Branch RAS, Ekaterinburg, Russia;*
[6]*Institute of Computational Mathematics RAS, Moscow, Russia*
e-mail: bazhan@vector.nsc.ru
**Corresponding author*

Key words: HIV-1, TAT-REV regulation of replication, feedback regulatory loop, mathematical modeling

Motivation and aim: The understanding of molecular genetic aspects of HIV-1 intracellular ontogenesis is a fundamental problem of biology. Progress in this area will contribute to the development of effective approaches in the treatment of AIDS. The aim of this work is to study the peculiarities of HIV-1 replication using a high-resolution mathematical model formulated with a system of ODEs.

Results: The developed mathematical model describes the regulation of HIV-1 replication through viral proteins Tat and Rev operating via a feedback mechanism, according to which the Tat protein is an activator, and Rev acts as repressor. The model describes: (1) transcription activation of proviral DNA by NF-kB cellular protein; (2) synthesis of short RNAs terminating in TAR-element; (3) Tat-specific transcription antitermination in TAR-element, resulting in synthesizing full-length genomic 9kb-RNA; (4) splicing 9kb-RNA in 4kb-RNA and 4kb-RNA in 2kb-RNA coding Tat and Rev proteins; (5) transport of 2kb-RNA from nucleus to cytoplasm; (6) cytoplasmatic synthesis of Tat and Rev proteins and their transport from cytoplasm to nucleus; (7) multiple binding of Rev protein to 9kb- and 4kb-RNA and their transport from nucleus to cytoplasm. The model also describes the degradation of proteins and RNA of all types both in the cytoplasm and in the nucleus. The model considers the dynamics of about 100 variables, characterizing the concentrations of specific substances: viral DNA, RNA, proteins and their complexes located in nucleus and cytoplasm. The values of the model parameters are taken from the literature or are prescribed on the basis of biological feasibility. Numerical analysis shows that at certain levels of transcription activity of proviral HIV-1 DNA integrated into the chromosome of a host cell, the dynamics of synthesis of virus-specific RNAs and proteins exhibits cyclic behavior. Numerical experiments have demonstrated that the cycle period is sufficiently long: ranging from 80 to 400 hours. The biological aspects of the oscillatory dynamics of HIV-1 replication are discussed.

Availability: Further development of the proposed mathematical model aims at revealing the key stages of HIV-1 replication which may be potential targets for the antiviral drug activity.

Acknowledgements: This work has been supported by the RAS Presidium Program "Fundamental science for medicine" and RFBR grants N 14-01-00477 and 14-04-91164.

BACKBONE ENCODINGS IN PROTEIN STABILITY PREDICTIONS

Liu J., Kang X.*
Life Science School, Hebei University, Baoding, 071002, China
e-mail: Liu@hbu.edu.cn; xjkang@hbu.edu.cn
**Corresponding author*

Key words: Protein Structure Function, Backbone Encodings, Support Vector Machines

Motivation and Aim: Single amino acid substitutions or site-directed mutagenesis in a protein chain can significantly change the protein thermodynamic stability. Amino acid residues around the mutation point within certain distance (<9.0 E, for example [1]) were encoded as mutation point environments into machine learning input information. The backbones of the protein structure, however, are crucial in maintaining protein tertiary structure and keeping protein stability.

Methods and Algorithms: The dataset DB3D_Sep05 [1], PopMuSic2.0 [2] were used as model building dataset, and D350 was used as independent validation test set. The LibSVM package was used for SVM training and prediction. **Protein Structure Backbone Representations**: The three longest segments of α-helices and β-sheets were encoded respectively for each protein molecule. For the overall three longest sustaining segments of helices or sheets, there are eight possible arrangements: HHH, EEE, HEH, EHE, HHE, EEH, HEE, and EHH. We represented the permutation pattern with 1 through 8 distinct numbers.

Results: The backbone encoding scheme improved more than 4% in overall prediction accuracies with 20-fold cross-validation tests. At least 5% improvement was observed with independent validation tests.

Conclusion: The backbone encoding scheme produce uniform data format from any protein molecules with no restrictions. The backbone encoding method can be used in other protein function related predictions.

References:

1. Capriotti E., Fariselli P. and Casadio R. (2005) I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res.*, **33**: W306-W310.
2. Dehouck Y., Kwasigroch J.M., Gilis D., and Rooman M. (2009) Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. *Bioinformatics*, **19**: 2537-2543.

THE TWO HYPOTHESES OF BAIKAL ENDEMIC SPONGE (LUBOMIRSKIIDAE) EVOLUTION

Maikova O.O.*, Sherbakov D.Yu., Belikov S.I.
Limnological Institute SB RAS, Irkutsk, Russia
e-mail: Maikova@lin.irk.ru
**Corresponding author*

Motivation and Aim: In order to resolve the timing of the speciation within the Lubomirskiidae family, we describe here the results of a comparison between one newly sequenced and previously described mitochondrial genomes of the endemic Baikalian (Lavrov 2010; Lavrov et al. 2012) and cosmopolitan sponges. The aim was to reconstruct the evolutionary history of the Lubomirskiidae, so that we could test two alternative hypotheses: one was a possible divergence of Lubomirskiidae about 25-30 Ma based on the paleontological data obtained by Martinson (1940) and the other a divergence of Baikalian endemic sponges in the Miocene, based on the presence of Lubomirskiidae spicules in Baikal bottom sediment aged about 10 million years.

Methods and Algorithms: For the phylogenetic analysis, individual genes were aligned with BioEdit using default parameters. Bayesian reconstructions were performed using MrBayes v. 3.2.1. The estimated parameters (GTR+G for cox2, atp6, nad6 and nad3; GTR+I+G for cox3, cob, atp9, nad4, cox1, nad2 and nad5; HKY+I for atp8; «mixed» parameter for nad4L and nad1) were used in Bayesian inference. To calibrate the phylogenetic trees, we set the values of substitution rates. To test the evolutionary hypothesis, we computed the marginal likelihoods ratio (Bayes factor) by Tracer v. 1.5 and used the stepping-stone method in Bayesian analysis.

Results: We determined the nearly complete mtDNA sequence of the Lake Baikal sponge *Baikalospongia bacillifera* (KJ192328) and analyzed it with those of a Baikalian sponge *L. baicalensis* (Lavrov, 2010), *R. echinata*, *B. intermedia profundalis, S. papyracea* (Lavrov et al. 2012) and a cosmopolitan freshwater sponge *E. muelleri* (Lavrov et. al., 2005).

In the first scenario of Baikalian sponge speciation, the origination and diversification of the Fam. Lubomirskiidae is assumed to have taken place in the Late Oligocene (Martinson 1940), with the common ancestor of species sequenced to date appearing some time between 4.3 to 11.7 Ma (expected 7.6). The divergence within the clades A and B took place between 0.5 to 5.1 Ma (expected 2.5) and 0.7 to 5.2 Ma (expected 2.7) respectively. The second scenario is based on palaeontological data using the presence of spicules (belonging to the 4 recent species) in bottom sediments dated to 10 Ma. If we assume this date as the upper limit for the Lubomirskiidae diversification, the common ancestor of Baikalian species should have appeared approximately at the border between the Pleistocene and Pliocene ca. 2.3 Ma (HPD=1.3 to 3.6 Ma), while *L. baicalensis, R. echinata, B. intermedia profundalis* and *B. bacillifera* diverged in the Pleistocene about 0,7 Ma, 95% HPD ranging from 0.15 to 1.5 Ma.

Both hypotheses were compared using a Bayes factor. The resulting log*B* was 4.737 in favour of the younger origin of the Lubomirskiidae. For the hypothesis of Lubomirskiidae divergence about 25-30 Ma, the geometric average of marginal log likelihood is -17162.61 and for the hypothesis of divergence at least 10 Ma is -17146.82. Therefore, the Bayes ratio test suggests strong evidence in favor of a divergence of the Lubomirskiidae closer to 10 Ma rather then to 30 Ma.

Conclusion: The origin of the Baikalian endemic sponge family Lubomisrskiidae, and further speciation within it, have been studied by comparing full sets of protein-coding sequences of the mitochondrial genome. It was confirmed that the common ancestor of all contemporary Lubomisrskiidae species diverged from a representative of the freshwater Spongillidae adapted to relatively warm water conditions. All subsequent evolutionary events may be explained by niche splitting along temperature, depth and possibly feeding preferences. Each of these events took place when global environmental fluctuations were the most favourable.

Availability: All programs are available over the Internet: BioEdit (http://bioedit.software.informer.com/7.1/), MrBayes v. 3.2.1. (http://mrbayes.sourceforge.net/download.php), Tracer v. 1.5 (http://tree.bio.ed.ac.uk/software/tracer/).

References:

1. Lavrov D.V., Maikova O.O., Pett W., Belikov S.I. (2012) Small inverted repeats drive mitochondrial genome evolution in Lake Baikal sponges. Gene 505:91–99.
2. D.V. Lavrov. (2010) Rapid proliferation of repetitive palindromic elements in mtDNA of the endemic Baikalian sponge Lubomirskia baicalensis, *Mol. Boil. Evol.,* 27(4): 757–760.
3. Lavrov D.V., Forget L., Kelly M., Lang B.F. (2005) Mitochondrial Genomes of Two Demosponges Provide Insights into an Early Stage of Animal Evolution. Mol Biol Evol 22:1231–1239.

TRANSPOSONS VS GENES: SURVIVAL STRATEGIES

Matvienko V.F.[1], Babenko V.N.*[2]
[1]*Novosibirsk State University, Novosibirsk, Russia;*
[2]*Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia*
e-mail: bob@bionet.nsc.ru
**Corresponding author*

Key words: retroposons, chromatin state, gene orientation

Motivation and Aim: One of the absolutely advantageous tools in *D.melanogaster* community is transposon screening system which allows studying various genomic segments. Recent advances proposed analogous PiggyBac screening system for human. In this paper we analyzed the distribution of both total and manually curated *P*-element sets with estimated mini-*white* marker gene suppression rate located in intergenic regions. The regions were classified based on mutual orientation of flanking genes representing 'head to head', 'tail to tail' and 'collinear' categories. The drastic difference in transposon density was revealed between categories: very few insertions were located in 'tail to tail' (convergent) intergenic regions. On the contrary, 'head to head' (divergent) intergenic regions harbor about a half of intergenic *P*-elements. In addition, considering mini-*white* marker gene suppression rate, the most frequent suppression was observed in convergent intergenic regions. Further analysis revealed chromatin specifics in each intergenic category. The work underlines non-random correlation of chromatin landscape in intergenic regions and mutual gene orientation at the flanks.

Methods and Algorithms: We use Flybase v.5 (2013/06) for retrieving gene locations and compiling protein coding gene set. We defined intergenic region as a region within neighbor genes not less than 2 bp. We considered euchromatic regions of chromosomes X, 1, 2, 3. The compilation resulted in 12426 intergenic regions. We segregated intergenic regions based on flanking genes mutual orientation. The statistics on the observed categories are listed in Table 1.

Sample of 2852 *P*-elements with mini-*white* marker gene we retrieved from (Babenko et al., 2010). Pscreen data set of 7616 *P*-element coordinates has been retrieved from 'pscreen' table at ucsc database (www.genome.ucsc.edu).

Results: The initial distribution of categorized intergenic regions is presented in Table 1. Elaborating the promoter number column, if we consider a gene pair oriented 'head to head'(common 5' region) we get two promoters in corresponding intergenic region (sometimes it is called a bi-directional promoter). Conversely, tail to tail genes yield no promoters in attributed intergenic region. Collinear genes harbor exactly one promoter in the intergenic region, correspondingly. As it could be seen from Table 1, the number of gene pairs maintaining convergent and divergent orientation is a bit abundant compared to collinear genes.

We calculated the number of intergenic insertions from each category for 3214 (from 7616 total) pscreen database and 1271 (from 2852 total) phenotyped insertions (see Materials and Methods) listed in Table 2. When plotted the categorized insertion numbers against the number of promoters we observed good linear regression approximation (Fig. 1). We should note that regression stays valid on small and long intergenic distances.

Lastly, we assessed the chromatin state specifics (Fillion et al., 2010) at the sites of intergenic insertions. We based on five chromatin states in the order of its activity: "red", "yellow", "green", "blue" and "black" that were derived in (Fillion et al., 2010). We excluded "green" one due to its small size and a deficit in insertion number (less than 10 total). The histogram of relative occurrence of chromatin in each intergenic category is presented on Fig. 2. It could be seen that most active chromatin is attributed to divergent intergenic areas. Most heterochromatic were convergent intergenic areas. It's worth noting that a significant part of insertions is located in Polycomb G (PcG) chromatin in each intergenic category ("blue", Fig. 2).

Conclusion: The study provides insights into the insertion strategy of various retroposon species both in insect and human based on their distribution specifics related to chromatin environment properties and a gene landscape. Possible synergistic outcome and a compromise evolutionary model are discussed.

References:

1. Babenko V.N., Makunin I.V., Brusentsova I.V., Belyaeva E.S., Maksimov D.A., Belyakin S.N., Maroy P., Vasil'eva L.A., Zhimulev I.F. 2010. Paucity and preferential suppression of transgenes in late replication domains of the D. melanogaster genome. BMC Genomics. 11:318.
2. Filion G.J., van Bemmel J.G., Braunschweig U., Talhout W., Kind J., Ward L.D., Brugman W., de Castro I.J., Kerkhoven R.M., Bussemaker H.J., van Steensel B. 2010. Systematic protein location mapping reveals five principal chromatin types in Drosophila cells. Cell 143:212-224.

COMPUTATIONAL INVESTIGATION OF HIGH PRESSURE AND TEMPERATURE INFLUENCE ON ARCHAEA PYROCOCCUS GENUS NIP7 PROTEIN STRUCTURE

Medvedev K.E.[1]*, Afonnikov D.A.[1, 2]
[1]*Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia;*
[2]*Novosibirsk State University, Novosibirsk, Russia*
e-mail: kirill-medvedev@yandex.ru
**Corresponding author*

Key words: Nip7 protein, high pressure, high temperature, molecular dynamic simulation, hyperthermophilic archaea

Motivation and Aim: High temperatures and pressures cause damage to living cells. There exist organisms, which colonize habitats extreme, life-incompatible for humans. Such conditions are near deep hot springs colonized by communities of organisms, the extremophiles. The mechanisms by which cell survival is provided are not well understood. Their elucidation would provide answers to some fundamental questions on the origins of life and the early adaptation of microorganisms, also on adaptation to the conditions of diverse ecological niches To investigate mechanisms of protein structure adaptation to high pressure environment we compare the dynamic properties of Nip7 proteins from hyperthermophilic archaea, *Pyrococcus abyssi* (deep sea habitat) and *Pyrococcus furiosus* (shallow water habitat) at different pressures (0.1-100 MPa) and temperatures (300, 373K).

Methods and Algorithms: MD simulations and structure analysis were performed using GROMACS [1].

Results: Obtained data suggested that the RNA-binding domain of the *P. abyssi* Nip7 protein is more resistant to the effects of high pressure. Moreover, analysis of computer models of Nip7 proteins from *P. abyssi* and *P. furiosus* showed that the solvent-accessible surface area of proteins decreases with the pressure increasing. At high pressures and temperatures, the *P. abyssi* protein model proves to be more compact and structurally more similar to the crystal model compared to the *P. furiosus* protein model under the same conditions. However, at high temperatures and atmospheric pressure, on the contrary, deep-water protein model demonstrates significantly greater fluctuations in the polypeptide chain than shallow-water model. This may be a consequence of the presence of a number of substitutions, which increase the hydrophobic fraction of the solvent accessible area of the surface of Nip7 protein in *P. furiosus* thereby promoting hydrophobic interaction at high temperatures and atmospheric pressures. The current analysis demonstrated the complex finely-tuned interplay between the effects of elevated pressures/temperatures that affect the dynamics of the Nip7 protein. This helped us to understand better the dynamic and evolutionary properties related to Nip7 protein functions.

Conclusion: In general, these data are consistent with the importance of hydrophobic interactions for the protein globule formation and the presumable mechanism of destruction of protein structures under high pressure.

References:

1. D.Van der Spoel et al. (2005) GROMACS: Fast, Flexible and Free. J. Comp. Chem. 26:1701-1718.

APOPTOTIC NUCLEAR VOLUME DECREASE: ANALYSIS OF CONFOCAL IMAGES AND MATHEMATICAL MODEL

Mikhaelis I.M.*, Chernyshev A.V., Yurkin M.A., Nekrasov V.M., Maltsev V.P.
Institute of Chemical Kinetics and Combustion, Novosibirsk, Russia;
Novosibirsk State University, Novosibirsk, Russia
e-mail: mi93work@mail.ru
**Corresponding author*

Key words: Confocal microscopy, apoptosis, HepG2, apoptotic ring, nuclear volume decrease

Motivation and aim: Apoptosis, a programmed cell death, is a safe mechanism for destroying damaged or useless cells without inflammation. The understanding of the apoptosis mechanism will help to diagnose and treat such diseases as: cancer, leukemia, lymphoma, immunodeficiency, infectious and virus disease. Unfortunately, there is no model, describing this process.

The first stage of the apoptosis is determined by the enzyme, PI3K, which enters into the nucleus and phosphorylates the histone proteins(H2) [1]. This results in chromatin condensation and the formation of the apoptotic ring. So, the aim is to develop molecular kinetic model based on this literature fact and the experimental data and to obtain several parameters relevant for medical and biological applications.

Methods and algorithms: We used the scanning laser confocal microscope to obtain stack of nuclear slices of tumor liver cells, called HepG2. We used two methods to measure a nuclear volume. First one uses the Sobel method for computing the gradient of intensity along the vector (from inside the nucleus to outside). If the value of a gradient changes rapidly, this pixel is assigned to a boundary surface. Everything inside this boundary is a nucleus; its volume is computed by summing enclosed areas in each layer. Overall, this approach determines the volume of star-shaped hull of the nucleus. Second method, software Imaris, is based on 3D interpolation, thus measuring only the volume occupied by the chromatin. Kinetic model is based on known five reactions between PI3K and related molecules and the osmotic balance.

Results: The developed model was used to fit the experimental data of nuclear volume decrease during the apoptosis. Several relevant model parameters were obtained with good accuracy, in particular the lag phase time, volume fraction of the chromatin in the nucleus, and the speed of the apoptosis.

Conclusion: The proposed model explains the formation and evolution of the apoptotic ring in the nucleus. The developed approach allows one to determine the dynamic characteristics of early apoptosis by experimental data. The total depletion of the nuclear volume agrees with the literature data.

References:

1. Solier S., Kohn K.W., Scroggins B., Xu W., Trepel.J., Neckers L., Pommier Y. Proc Natl Acad Sci USA. (2012);109(32):12866-72. doi: 10.1073/pnas.1203617109.

ASSOCIATION BETWEEN microRNA AND UTRS FROM HUMAN TLR GENES

Mitra C.K.*, Meena A.K.
School of Life Sciences, University of Hyderabad, Hyderabad, India 500046
e-mail: c_mitra@yahoo.com
**Corresponding author*

Key words: 3'-UTR, 5'-UTR, microRNA, TLR gene, miRNA

Motivation and Aim: To find common sequences between microRNA (miRNA) and 3'-UTR or 5'-UTR sequences of 10 TLR genes present in human.

Toll Like Receptors (TLRs) are a class of proteins that play key roles in innate immune system. Most TLR genes have multiple transcripts which contain 3'-UTR and 5'-UTR (UnTranslated Regions) sequences. Regulatory region within UTRs can influence polyadenylation, translation efficiency, localization and stability of the product mRNA. TLR genes were selected for this study because (i) they are reported to be primitive in nature, (ii) they are closely related in both structure and function, and (iii) they are believed to be relatively conserved during evolution. miRNAs are small non-coding RNA molecules which play important roles in transcriptional and post-transcriptional gene expression. In human ~1800 miRNAs have been identified so far.

Algorithms: UTRs of 10 human TLR genes were downloaded from www.ensembl.org. Sequence data for complete mature miRNAs sequence present in human were downloaded from www.mirbase.org. All sequences were saved in a file. In this file, U was replaced by T for ease of comparison.

To find out common unique sequence matches between miRNA and 3'-UTR (and 5'-UTR). A C program was written to search for all possible 6 nucleotides (6nt) sequences in data sets. 6nt sequences were seen on terminal with respective frequency. The 6nt sequences of 3'-UTR and 5'-UTR were taken and extending these 6nt sequences upto 12nt sequences. For unique sequence matches 6nt sequences of 3'-UTR and 5'-UTR were searched in miRNA file and this 6nt sequences extending till no further matches were found. 'grep', 'sort', 'uniq' commands were extensively used to locate and filter the matching sequences.

Results: The numbers of unique sequence matches (6nt-12nt) present in (3'-UTR &5'-UTR) and 6nt -12nt miRNA are different. The number of unique sequence matches reduces from 6nt to 12nt . Only 40 miRNAs share 12nt sequence similarity with 3'-UTR of 10 human TLR genes. About 18 miRNAs share 12nt sequence similarity with 5'-UTR of 10 human TLR genes.

Table 1: No. of common unique sequences matches

Common sequences between	6nt	7nt	8nt	9nt	10nt	11nt	12nt
miRNA and 3'-UTR	3079	5427	5427	1438	432	120	40
miRNA and 5'-UTR	2809	4039	2736	1088	333	101	18

Conclusion: Number of 12nt sequences matching between 3'-UTR and miRNA is 40. Similarly, number of 12nt sequences common between 5'-UTR and miRNA is 18. These sequences may have some regulatory role in the regulation of TLR transcripts.

EVIDENCE FOR EXTENSIVE NUCLEOSOME CROWDING IN YEAST CHROMATIN

Morozov A.V.
Department of Physics & Astronomy and BioMaPS Institute for Quantitative Biology, Rutgers University, USA

Eukaryotic genomes are organized into arrays of nucleosomes. Each nucleosome consists of up to 147 base pairs (bp) of genomic DNA wrapped around a histone octamer core. The resulting complex of DNA with histones and other regulatory and structural proteins forms a multi-scale structure called chromatin. At the most fundamental level of chromatin organization, arrays of nucleosomes form 10-nm fibers which resemble beads on a string, and in turn fold into higher-order structures. Depending on the organism and cell type, 75-90% of genomic DNA is packaged into nucleosomes. Nucleosomal DNA may transiently peel off the histone octamer surface due to thermal fluctuations or interactions with chromatin remodelers. Thus neighboring nucleosomes may invade each other's territories through DNA unwrapping and translocation, or through initial assembly in partially wrapped states. A recent high-resolution map of distances between dyads of neighboring nucleosomes in baker's yeast (*S. cerevisiae*) has revealed that nucleosomes frequently overlap DNA territories of their neighbors. Indeed, at least 25% of all yeast nucleosomes are separated by less than 147 bp, implying that at least one nucleosome in the pair is partially unwrapped. This conclusion is supported by lower-resolution maps of *S. cerevisiae* nucleosome lengths based on micrococcal nuclease digestion and paired-end sequencing. The average length of wrapped DNA follows a stereotypical pattern in genes and promoters: nucleosomes tend to be more unwrapped in promoters and less unwrapped in coding regions.

To explain these observations, we have developed a biophysical model which employs a 10-11 bp periodic histone-DNA binding energy profile. The profile is based on the pattern of histone-DNA contacts in nucleosome crystal structures, and the idea of linker length discretization caused by higher-order chromatin structure. Our model is in agreement with the observed genome-wide distributions of inter-dyad distances, wrapped DNA lengths, and nucleosome occupancies. Furthermore, our approach explains *in vitro* measurements of accessibility of nucleosome-covered target sites, and nucleosome-induced cooperativity between DNA-binding factors. We rule out several previously proposed scenarios of histone-DNA interactions as inconsistent with the genomic data. The extent of nucleosome crowding in the yeast genome suggests that its treatment should be included in all future models of nucleosome positioning and energetics.

TRANSCRIPTOMICS ANALYSIS OF *DROSOPHILA MELANOGASTER* AGING

Moskalev A.A.[1, 2, 3], Plyusnina E.N.[1, 2], Peregudova D.O.[1], Shaposhnikov M.V.[1, 2], Snezhkina A.[4], Kudryavtseva A.V.[4], Baranova A.V.[5]
[1]*Institute of Biology of Komi Science Center Ural Branch RAS, Syktyvkar, Russia;*
[2]*Syktyvkar State University, Syktyvkar, Russia;*
[3]*Moscow Institute of Physics and Technology (State University), Dolgoprudny, Russia;*
[4]*Engelhardt Institute of Molecular Biology RAS, Moscow, Russia;*
[5]*Research Center for Medical Genetics RAMS*

Drosophila melanogaster is very convenient model in aging study because it is well-researched and short-living. In our study we applied different bioinformatics approaches to public available and own transcriptomics data to obtain sex-independent reproducible biomarkers of *Drosophila* aging. Differentially expressed genes were also underwent pathway analysis by GO-BP and KEGG. In this work we discuss similarity of *Drosophila's* and human's transcriptomics biomarker of aging.

This work was supported by RFBR grant N 14-04-01596 and the grant of the President of Russian Federation MD-1090.2014.4.

ALLELIC COADAPTATION AND FITNESS LANDSCAPE PREDETERMINE THE OPTIMAL EVOLUTIONARY MODE IN PROKARYOTIC COMMUNITIES: A SIMULATION STUDY

Mustafin Z.S., Matushkin Yu.G., Lashin S.A.*
Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia;
Novosibirsk State University, Novosibirsk, Russia
e-mail: lashin@bionet.nsc.ru
**Corresponding author*

Key words: evolution, simulation, mathematical model, multiscale model, coadaptation, allelic combination, coevolution

Motivation and Aim: Impact of genes on organism phenotype is not one-to-one. On the one hand, each trait is determined by a set of genes (polymery). On the orher hand, each gene affects the expression of several phenotypic traits (pleiotropy). Pleiotropy and polymery reflect genetic interactions. Thereby, external manifestations of each gene depends on its genetic context – gene network structure. Genetic processes cause the origin of novel allelic combinations, most optimal of which fix then in a population owing to the natural selection. Mutation in a gene (i.e. origin of a novel allele) included in a gene network frequently causes disorders of its functioning. Such disorders can be fixed both by return to a previous allelic state or by the origin of novel allelic state of another gene in the same gene network. This phenomenon – allelic coadaptation, is increasingly studied in modern biological literature. We have performed simulations and analyzed the impact of allelic coadaptations and fitness landscapes on evolutionary modes in prokaryotic communities.

Methods and Algorithms: Simulations have been performed with software package "Haploid evolutionary constructor" [1]. We considered one-specie community of prokaryotic cells, which reproduction determined by efficiency of a non-specific substrate utilization. The efficiency in its turn determined by a two-gene gene network. Various allelic combinations gave different utilization efficiency, which described by a multimodal fitness function.

Results: We have revealed three parameters affecting the possibility of a population to reach the fittest state along with the part of the fittest subpopulation in the whole population. These parameters are fitness function landscape, mutation rate, and fitness bonus of the fittest cells.

Conclusion: It has been shown that the reach of coevolutionary success in coadaptive gene networks is determined by differences in fitness of various allelic combinations as well as by mutation rate. It has also been shown that different fitness landscapes have their own optimal frequencies of mutation occurrences. Moreover, in some cases the optimal evolutionary mode is gradualism, in other cases the optimal is the saltation evolution.

Availability: http://evol-constructor.bionet.nsc.ru

Acknowledgements: The work is supported by the RFBR grant 13-04-00620, the Stipend of the President of Russian Federation, the Program of the RAS Presidium № 28.

References:

1. Lashin S.A., Matushkin Yu.G., Suslov V.V., Kolchanov N.A. Computer modeling of genome complexity variation trends in prokaryotic communities under varying habitat conditions // Ecological Modelling, 2012, Vol. 224, No.1, pp. 124-129.

SLOWDOWN OF GROWTH ACTS AS A SIGNAL TRIGGERING CELLULAR DIFFERENTIATION

Narula J.[1#], Kuchina A.[2#], Süel G.M.[2], Igoshin O.A.[1*]
[1]*Department of Bioengineering, Rice University, Houston, TX, USA;*
[2]*Division of Biological Sciences, UCSD, San Diego, CA, USA*
[#]*co-first authors*
**Corresponding and presenting author*

Key words: *B. subtilis*, sporulation, starvation

Motivation and Aim: Starving *B. subtilis* cells cease vegetative growth and execute a multistage gene-expression program resulting in formation of stress-resistant spores [1]. This program is triggered by activation the sporulation master regulator, Spo0A (0A) that controls the expression of hundreds of genes. Previously we showed that a threshold level of 0A activity commits cells to sporulation [2]. However it is unclear how the diverse set of environmental and metabolic signals can be simultaneously detected and properly encoded into 0A activity. Recent studies of sporulation raised further questions by revealing cells respond to starvation by producing a pulse of 0A activity every cell-cycle [3].

Methods and Algorithms: Here we use a combination of mathematical modeling, single-cell microscopy and synthetic biology to uncover the mechanisms underlying sensing starvation, the pulsatile response of the sporulation phosphorelay and delayed commitment of cells to sporulation.

Results: Using a mathematical model of the phosphorelay, we proposed that the chromosomal positions of two genes controlling the activation of the sporulation master regulator 0A, *kinA* and *0F*, play the pivotal role in coupling sporulation to the cell-cycle. Since *0F* located close to the origin and *kinA* near the terminus of the *B. subtilis* chromosome, each DNA replication cycle leads to a transient imbalance in their gene dosages. This imbalance is amplified by feedback loops in the sporulation network and thereby triggers pulses of 0A activity once per cell-cycle. Our model also predicts that as growth slows down during starvation, pulses of 0A activity increase in amplitude. Thus the cell growth rate can act as the starvation signal and a threshold of 0A activity acts as a growth-rate threshold. Analysis of single-cell data confirms predicted relationship between cell growth, 0A activity and sporulation.

Conclusion: We show that regulatory networks can be coupled to the cell-cycle through proper chromosomal arrangement of genes. Moreover using the cell growth rate as the starvation signal allows the cells to integrate nutritional information without specific metabolite sensing.

Acknowledgements: This work was supported by NSF grant MCB-1244135.

References:

1. Errington J. (1993) *Bacillus subtilis* sporulation: regulation of gene expression and control of morphogenesis. *Microbiol Rev.* **57**, 1-33.
2. Narula J., et al. (2012) Ultrasensitivity of the *Bacillus subtilis* sporulation decision. *PNAS* **109**, E3513-22.
3. Levine J.H., et al. (2012) Pulsed feedback defers cellular differentiation. *PLoS Biology* **10**, e1001252.

BIOINFORMATIC ANALYSIS OF ENDO-β-XYLANASES FROM *PLANCTOMYCETES*

Naumoff D.G.
Winogradsky Institute of Microbiology RAS, Moscow, Russia
e-mail: daniil_naumoff@yahoo.com

Key words: glycoside hydrolase, endo-β-xylanase, Planctomycetes, protein evolution, lateral gene transfer, molecular phylogeny, paralogue, orthologue, GH10, CAZy

Motivation and Aim: The endo-β-xylanase (EC 3.2.1.8 & EC 3.2.1.32) is a widespread and industrially important enzyme which is responsible for degradation of heteroxylans. The catalytic domains of currently known enzymes of this group belong to 13 different glycoside hydrolase families. Among them there are GH10 and GH11 families, which do not contain enzymes with any other enzymatic activities. The vast majority of all experimentally characterized endo-β-xylanases belong to these two families. *Planctomycetes* is a bacterial phylum with poorly characterized functional capabilities. According to the CAZy database (http://www.cazy.org/), 12 out of the 13 families either are encoded in a limited number of *Planctomycetes* genomes, or do not encoded in them at all. The GH10 family of glycoside hydrolases is the only one represented in almost all *Planctomycetes*. A computational analysis of a group of hypothetical GH10 endo-β-xylanases from the *Planctomycetes* is the purpose of this work.

Methods and Algorithms: Protein sequences were retrieved from the NCBI database. Multiple sequence alignment of GH10 domains was made in BioEdit program. The phylogenetic trees were built using programs of PHYLIP package.

Results: Iterative screening of the database revealed 61 GH10-containing proteins from various *Planctomycetes*. Pairwise comparison allowed us to divide them into two groups based on the sequence similarity. One of them is represented by a single protein in every heterotrophic *Planctomycetes*. These proteins form a phylogenetically distinct and tight cluster. This cluster is highly congruent with the tree constructed for the 16S rRNA genes, thus suggesting the vertical transfer of the corresponding endo-β-xylanase genes and their functional importance in *Planctomycetes*. The second group is present only in three genera of the *Planctomycetes*: *Gemmata*, *Phycisphaera*, and *Rhodopirellula*. Usually it is represented by several paralogues. The number of paralogues could be different even in strains of the same species, as it has been shown for *Rhodopirellula baltica* and '*R. europaea*'. Proteins of the *Planctomycetes* from this group are not always the closest homologues for each other. The phylogenetic analysis of the proteins from the second group confirmed that the *Planctomycetes* proteins do not form a distinct cluster. These data suggest evolution of genes encoding these proteins by multiple cases of independent horizontal gene transfers to *Planctomycetes* genomes.

Conclusion: *Planctomycetes* members of the GH10 family of glycoside hydrolases (hypothetical endo-β-xylanases) belong to two distinct groups. One of them is globally distributed among heterotrophic *Planctomycetes* and evolves by the vertical transfer. The second group is quite rare and has appeared by several evolutionary recent independent horizontal gene transfers.

References:

1. D.G. Naumoff, A.A. Ivanova, and S.N. Dedysh (2014) Phylogeny of β-xylanases from Planctomycetes, *Molecular Biology* (Moscow), **48** (3), *in press*.

DE NOVO ASSEMBLY OF THE MITOCHONDRIAL GENOME OF ~5000-YEAR-OLD HUMAN FROM NORTH CAUCASUS

Nedoluzhko A.V.[1]*, Boulygina E.S.[1], Sokolov A.S.[1], Tsygankova S.V.[1], Schubert M.[2], Gruzdeva N.M.[1], Rezepkin A.D.[4], Orlando L.[2], Prokhortchouk E.B.[1, 3]*

[1]*National Research Center «Kurchatov Institute», Moscow, Russia;*

[2]*Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Copenhagen, Denmark;*

[3]*Center of Bioengineering, Russian Academy of Sciences, Moscow, Russia;*

[4]*Institute for the History of Material Culture, Russian Academy of Sciences, Saint Petersburg, Russia*

e-mail: nedoluzhko@gmail.com, prokhortchouk@gmail.com

**Corresponding author*

Key words: Novosvobodnaya culture, Maykop culture, haplogroup, mitochondrial DNA, sequencing, genomics

Motivation and Aim: The Novosvobodnaya culture is known as a Bronze Age archaeological culture in the Western Caucasus region of Southern Russia. It is dated to the end of 4th millennium B.C. and seems to partially coincide with the Maikop culture that resulted in two hypotheses of the Novosvobodnaya culture origin. One hypothesis suggests that the Novosvobodnaya culture is a phase of the Maykop culture, whereas another one considers it as an independent event based on specific differences in material culture items found in the graves. A comparison of Novosvobodnaya pottery to Funnelbeaker (TRB) pottery from Central Europe allowed suggestion that the Novosvobodnaya culture was developed under the influence of Indo-European culture. Nevertheless, the Novosvobodnaya culture origin is still debated.

Methods and Algorithms: Ancient DNA was extracted from ~5000-year-old human remains from Klady kurgan grave at Novosvobodnaya (now the Republic of Adygea, Russia). DNA-libraries were created and 58771105 of reads were generated using Illumina GAIIx with on average 13.4x mitochondrial (mt) DNA coverage.

Results: Here we applied the Next-Generation Sequencing to study ~5000-year-old human remains from Klady kurgan grave at Novosvobodnaya (now the Republic of Adygea, Russia). Using our sequencing data we postulate that Novosvobodnaya human remains has V7 mtDNA haplogroup.

Conclusion: mtDNA haplogroup was defined as V7 suggesting an impact of TRB culture in the Novosvobodnaya culture development and supporting a model of the connection between Novosvobodnaya and early Indo-European cultures.

SARP: AN ALGORITHM FOR ANNOTATION OF THE COMPOSITIONALLY BIASED REGIONS IN PROTEIN SEQUENCES

Nizhnikov A.A.*, Antonets K.S., Inge-Vechtomov S.G.
St. Petersburg State University, St. Petersburg, Russia;
St. Petersburg Branch of N.I. Vavilov Institute of General Genetics RAS, Russia
e-mail: ant.nizhnikov@gmail.com
**Corresponding author*

Motivation and Aim: The composition of subunits (nucleotides, amino acids) is one of the key features of biological sequences. Compositional biases are local shifts in amino acid or nucleotide frequencies that can occur as an adaptation of an organism to an extreme ecological niche or as the signature of a specific function or localization of the corresponding protein. Previously, a method for annotating compositional bias called LPS (Lowest Probability Subsequences) was described [1]. It is based on the calculation of probabilities for defined regions in protein sequences, thus providing accurate detection of biased subsequences. Recently, we developed SARP (Sequence Analysis based on the Ranking of Probabilities), a novel algorithm for the annotation of compositional biases, which is based on the original method of ranking the subsequences by their probabilities [2]. The goal of the study is to compare efficiencies of SARP and traditional LPS algorithm.

Methods and Algorithms: The detailed descriptions of SARP and LPS algorithms are given in [2] and [1], correspondingly. The probability threshold in all cases was set as 10^{-12}. The generation of the set of random 1000 proteins of yeast *Saccharomyces cerevisiae*, was performed by selecting of the top proteins from the list of proteins sorted by accession.

Results: We performed the benchmarking of SARP with the original LPS algorithm using 1000 random protein sequences from *Saccharomyces cerevisiae*. For this set, the average length of a protein sequence was 489.2 residues, and the total length 489,217 residues. The running time for SARP was 1465 seconds, while the same for LPS was 341,914 seconds. The accuracy of SARP was demonstrated to be identical with LPS algorithm, with the exception of one very long compositionally biased region (GI 6319255), whose boundaries detected by SARP were very close to the boundaries found by LPS.

Conclusion: SARP provides the same accuracy as the previously published LPS algorithm but performs at an approximately 230-fold faster rate. It is recommended for use when working with large datasets to significantly reduce the time and resources required.

Availability: The source code for SARP is available upon e-mail request.

References:

1. P.M. Harrison, M. Gerstein. (2003) A method to assess compositional bias in biological sequences and its application to prion-like glutamine/asparagine-rich domains in eukaryotic proteomes, Genome Biol., 4: R40.
2. K.S. Antonets, A.A. Nizhnikov. (2013) SARP: a novel algorithm to assess compositional biases in protein sequences, Evol. Bioinform., 9: 263-273.

COMPUTATION MODELING OF VASCULAR PATTERNING IN PLANT ROOTS

Novoselova E.S.*, Mironova V.V., Kazantsev F.V., Omelyanchuk N.A., Likhoshvai V.A.
Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
e-mail: novoselova@bionet.nsc.ru
**Corresponding author*

Key words: auxin, vascular tissue, auxin carriers, patterning, mathematical model

Motivation and Aim: The plant vascular system has species-specific determined vascular pattern – relative localization a certain amount of phloem and xylem elements in central cylinder. It has been shown experimentally that auxin is key regulator of procambium differentiation to xylem and phloem. Unevenly distributed in the tissues auxin acts like morfogen and direction of the cell differentiation is given by the local auxin concentration (Aloni, 2011). Expression of auxin carriers PIN and AUX is regulated by auxin concentration (Vieten et al., 2005). However precise mechanisms of auxin impact on the vascular system organization are still poorly understood. There is a need in study this mechanisms using systems biology.

Methods: Here we present a 2D model for investigation auxin distribution on the cross section of the root central cylinder. The processes of auxin transporters synthesis, diffusion, degradation and active transport described on the basis of the mass action law and in terms of generalized Hill functions (Likhoshvai, Ratushny, 2007, Mironova et al., 2010, Mironova et al., 2012). The model has been assembled using the MGSModeller software. Numerical simulations and visualization of calculation were carried out using Matlab. The values of parameters were adjusted so that the pattern of auxin and auxin carriers expression in the root central cylinder obtains different types of the symmetry.
Results: Results of simulations have highlighted possible key features of auxin influence on the vascular patterning. It has been suggested that intensity of auxin incoming from the shoot to the root vascular cells and regulation of auxin carriers synthesis are important morphogenetic factors that ensure the formation of different types of root vascular system.

Acknowledgements: The work is partially supported by the RFBR grants 13-01-00344, RAS program “Molecular and cellular biology” (project 6.6), Integration project SB RAS 80 and budget project (VI.61.1.2).

FEATURES 8-oxo-dGTP BEHAVIOR IN ACTIVE SITE OF HUMAN DNA POLYMERASE β: MOLECULAR DYNAMICS STUDIES

Nyporko A.Yu.
Taras Shevchenko National University of Kyiv, Kyiv, Ukraine
e-mails: dfnalex@gmail.com, anyporko@univ.net.ua

Key words: 8-oxo-dGTP, DNA polymerase β, molecular dynamics, structural analysis, delayed dissociation

Motivation and aim: The oxidized bases in the composition of DNA as well as DNA precursors (dNTPs) appearing in living cell as a result of oxidative stress are the one of major sources of genomic instability. Among oxidized bases, the 8-oxo-dG is the most ubiquitous. This compound has a high mutagenic potential due to its ability to preferably interact with adenine instead of cytosine. In particular, the 8odG in the composition of the incoming nucleotide triphosphate (8oxoGTP) is able to immediately incorporate into the growing DNA chain and, thus, to cause the invert replacement dA>dC because it is possible to pair with the incoming dCTP as well as dATP in the next round of DNA replication. The efficiency of 8oxodG incorporation in growing DNA clearly depends on the nature of appropriate DNA polymerases. One of the most sensitive to 8-oxo-dGTP is the eukaryotic DNA polymerase β (pol β). The binding of 8-oxo-dGTP in the pol β active center can result in two different molecular events. First of them is the incorporation of 8-oxo-G into a growing DNA chain, the other is a discrimination of 8-oxo-dGTP from the active center. While effects of incorporation of this modified guanine in DNA are well studied, the immediate consequences of 8-oxo-dGTP discrimination are still unclear. Thus, the objective of the present investigation is the structural analysis of the behavior of 8-oxo-dGTP molecule in (the area of) the active site of human DNA polymerase β using molecular dynamics (MD) calculation.

Methods: The studied system were prepared on the basis of the X-ray derived structure of ternary complex "human DNA polymerase β: DNA: 8-oxo-dGTP" at 2.0 A resolution deposited in RCSB Protein Data Bank (access code is 3MBY). Molecule of 8-oxo-dGTP in pol β active site is paired with adenine of the template strand. All MD computations were performed with GROMACS software version 4.6.5 using CHARMM27 force field with implemented CMAP. The topology of 8-oxo-dGTP was performed via web-based tool SwissParam. To simulate intracellular conditions all further calculations were carried out in water solution containing 0.15M. Four productive MD trajectories were calculated for 100 ns time interval at 300 K. Fluctuations of 8-oxo-dGTP structure in free and bound states were evaluated using the *rho* method with *g_rms* modules of the GROMACS software. All energy parameters of the investigated system and its components were calculated with *g_energy* module. The dynamics of hydrogen bonds was calculated using the *g_hbond* module.

Results: The principle phenomenon revealed as investigation result is existence of two cardinally different models of behavior inherent to 8-oxo-dGTP molecule. In two cases the ligands loses the connections with template dA and starts to migrate inside of enzyme space (migrate trajectories). In the other two cases 8-oxo-dGTP stably stays in DNA polymerase active site, "keeps in touch" with template nucleotide and maintains the hydrogen bonds with it (stable trajectories). The 8-oxo-dGTP spatial structure in stable trajectories appears to be sufficiently rigid despite the presence of number of bonds around which the free rotation is possible, and its conformational energy is characterized by high stability over the time of studied MD. Average values of energy (–10229.7 and –10227.1 kJ/mol) are practically the same for both cases. Amino acid microenvironment of 8-oxo-dGTP also practically doesn't change over the studied MD interval. Thus, stable variants of 8-oxo-dGTP behavior evidently correspond to case of the further incorporation modified 8-oxo-dG into growing DNA strand. The behavior of 8-oxo-dGTP molecule in migrate trajectories is significantly more complicated. The 8-oxo-dGTP loses the H-bonds with template dA6 (at 11 and 6.5 ns of MD in first and second case respectively) and starts to migrate in DNA polymerase space. The 8-oxo-dGTP spatial structure regularly exhibits much more flexibility in comparison to itself behavior in stable trajectories that reflects in corresponded values of individual atomic fluctuations. However, contrary to the expectations the general levels of conformational energy of 8-oxo-dGTP as well as energy fluctuation patterns in both migratory trajectories are completely time stable. The average values of conformational energy are -9938.6 and -10018.6 kJ/mol for trajectories 1 and 2 respectively that is slightly more than corresponded values for stable trajectories. The 8-oxo-dGTP movement pathways of don't coincide each other that is confirmed by differences of their conformational spaces and amino acid microenvironment.

Conclusions: It seems to be the most important that 8-oxo-dGTP not only doesn't leave the enzyme space but directly prevent transition of DNA polymerase from closed to open conformation as well as the further binding of incoming dNTP. This observation lets a possibility to consider it as natural inhibitor of DNA pol β activity and possible intracellular regulator which mediates the direct transition of the cell from normal state to programmed cell death omitting the malignancy stage.

Acknowledgements. Author thanks the Extreme Science and Engineering Discovery Environment (XSEDE) for the award allocation number TG-DMR110088.

INTERACTION BETWEEN miRNA AND mRNA OF MYB TRANSCRIPTIONAL FACTORS FAMILY GENES OF MAIZE

Orazova S.B.*, Sagaydak A.I., Ivashchenko A.T.
Al-Farabi Kazakh National University, Almaty, Kazakhstan
e-mail: saltanat.orazova@kaznu.kz
**Corresponding author*

Key words: miRNA, mRNA, binding site, MYB family, maize

Motivation and Aim: Transcriptional factors are important modulators of genes expression in plants genomes. Transcriptional factors of MYB family involved in regulating expression of genes, that determined the growth, development and plant responses to stress. The expression of MYB genes strongly depends on miRNA, which modulate the synthesis of proteins in the process of mRNA translation. The influence of miRNA to MYB gene expression of plants is poorly understood, so it is important to evaluate a binding degree of miRNA and mRNA of maize MYB genes.

Methods and Algorithms: The binding sites of 322 zma-miRNA of maize in mRNA of 216 gene of MYB family was revealed using the program MirTarget, which defines the location of the miRNA with mRNA 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 miRNA with mRNA selected by $\Delta G/\Delta G_m$ equal or above to 85%.

Results: For 12 miRNA of miR156a-1 family were detected only one target genes with the magnitude $\Delta G/\Delta G_m$ equal to 85-90%. miR156a*-k* were associated with mRNA from one to three MYB genes with the value $\Delta G/\Delta G_m$ equal to 85-88%. Consequently, miR156 family members can control the expression of several genes of MYB family. Each miRNA of miR159a-k family was associated with one site in mRNA of 11 MYB genes. The value $\Delta G/\Delta G_m$ for these interactions was equal to 85-93%. miR159e also bonded with mRNA of 11 genes, including mRNA of AC217264.3 and GRMZM2G070523 genes, which were fully complementary: the value $\Delta G/\Delta G_m$ was 100%. miRNA of miR159e* associated with 19 genes mRNA, including mRNA AC217264.3 gene, which was fully complementary. This phenomenon is unique because the binding sites of miR159e and miR159e* located in the CDS of mRNA. miR162* bonded with mRNA of 13 MYB genes with the value $\Delta G/\Delta G_m$ equal to 85-89%. miR164a* had binding sites in mRNA of 17 MYB genes and value $\Delta G/\Delta G_m$ of their interaction varied from 85 to 91%. miR166b-i and miR166a*-n* had several target genes. Each of miR167a*-j* had two or more of target genes of MYB family. Family members of miR169a-r, and miR169a*-p* had binding sites in mRNA of some MYB genes and value $\Delta G/\Delta G_m$ varied from 85 to 88%. miR171b*-m* had the binding sites in of some mRNA MYB genes. Each miRNA of miR319a-d and miR319a*, c* families had a binding sites in some mRNA MYB genes. Other miRNA had a smaller number of binding sites in mRNA of MYB genes, but eventually expression of 88 MYB family genes depended on miRNA of maize. In all cases, the binding sites of miRNA with MYB gene mRNA were in CDS of mRNA.

Conclusion: Localization of the binding sites in CDS of mRNA typical for MYB, SPL and GRF family genes of plants. Localization of binding sites in CDS allows to saving the regulation of transcription factors gene expression by miRNA during millions of years of the plants evolution.

ANALYSIS OF THE BINDING MODE OF ANTIVIRAL PEPTIDES TO THE FLAVIVIRUS ENVELOPE PROTEIN E BASED ON PROTEIN–PROTEIN DOCKING

Orlov A.A.[1, 2]*, Osolodkin D.I.[1, 2], Palyulin V.A.[1], Zefirov N.S.[1]
[1]*Department of Chemistry, Lomonosov Moscow State University, Moscow, Russia;*
[2]*Chumakov Institute of Poliomyelitis and Viral Encephalitides, Moscow, Russia*
e-mail: rnase@qsar.chem.msu.ru
**Corresponding author*

Key words: flaviviruses, antiviral peptides, envelope proteins, homology modeling, protein–protein docking

Motivation and Aim: The *Flaviviruses* genus consists of 70 viruses such as dengue virus, yellow fever virus, Japanese encephalitis virus, tick-borne encephalitis virus, etc. They cause such severe diseases as hemorrhagic fevers and encephalitises. There is no appropriate specific treatment of flavivirus infections, and vaccines exist only against yellow fever, Japanese encephalitis and tick-borne encephalitis. A promising method to inhibit reproduction of flaviviruses is the prevention of the membrane fusion between the virus and the host cell. Fusion of flaviviruses is mediated by a distinct dimer to trimer rearrangement of the envelope E proteins that occurs in low pH medium of the endosome. Thus, two possible targets for the inhibition of the fusion process are dimeric and trimeric forms of the envelope E protein. Most small molecule inhibitors of reproduction are considered to interact with the dimeric form of the E protein. On the contrary, stem-derived peptides are assumed to interact with the trimeric form of the E protein. The X-ray data structures for peptide-E protein complex are not available. The sequence–activity relationships are also unclear. Our aim was to clarify the existing relationships by means of protein–protein docking.

Methods and Algorithms: homology modeling was performed in Modeller 9.11, docking of peptides was performed on the PatchDock and FiberDock servers, analysis of hydrophobicity profile was carried out with Membrane Protein Explorer (MPEx).

Results: The peptide-protein docking allowed us to distinguish between active and inactive stem–derived peptides according to their docking scores. Putative binding sites for these peptides were found on the trimeric form of the E protein. Corresponding peptides from other flaviviruses were docked and the same distribution of docking scores was observed. New peptides were designed based on the analysis of Whimley – White hydrophobic profile of the E protein sequence and their ability to interact with the trimeric form of the E protein was evaluated.

Conclusion: Peptide-protein docking allowed us to identify possible binding sites of antiviral peptides on the surface of the trimeric form of the E protein. Potential antiflaviviral peptides were designed with the help of Whimley–White hydrophobic scale analysis.

Acknowledgments: This work was partially supported by Russian Foundation for Basic Research grant №14-03-31566.

EXPERIMENTALLY VERIFIED TRANSCRIPTION FACTOR BINDING SITES MODELS APPLIED FOR COMPUTATIONAL ANALYSIS OF CHIP-Seq DATA

Oshchepkov D.Y.[1]*, Levitsky V.G.[1, 2], Kulakovskiy I.V.[3, 4], Ershov N.I.[1], Makeev V.J.[3, 4], Merkulova T.I.[1, 2]

[1]*Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia;*
[2]*Novosibirsk State University, Novosibirsk, Russia;*
[3]*Engelhardt Institute of Molecular Biology RAS, Moscow, Russia;*
[4]*Department of Computational Systems Biology, Vavilov Institute of General Genetics RAS, Moscow, Russia*
e-mail: diman@bionet.nsc.ru
**Corresponding author*

Key words: ChIP-Seq, EMSA, Transcription factor binding sites, FoxA, SiteGA, PWM, Transcription factor binding model, Dinucleotide frequencies

Motivation and Aim: ChIP-Seq is widely used for detection of genomic segments bound by transcription factors (TF), either directly at DNA binding sites (BSs) or indirectly via other proteins. Nowadays different approaches to identify TFBSs within ChIP-Seq peaks are implemented. However, the absence of direct experimental verification of these models makes it difficult to set a threshold to avoid recognition of too many false-positive BSs and to compare the actual performance of the models, making it complicated to use these models for the ChIP-Seq data interpretation.

Methods and Algorithms: Four computational FoxA BS prediction models of two fundamental classes: pattern matching based on existing training set of experimentally confirmed TFBSs (oPWM and SiteGA) and de novo motif discovery (ChIPMunk and diChIPMunk) were implemented and compared for analysis of ChIP-Seq data for FoxA2 binding loci in mouse adult liver and human HepG2 cells. We experimentally evaluated affinity of 64 predicted FoxA BSs using EMSA that allows distinguishing sequences able to bind TF to properly select prediction thresholds for the models.

Results: Thousands of reliable FoxA BSs within ChIP-Seq loci from mouse liver and human HepG2 cells were identified as a result. We also found that the performance of conventional position weight matrix (PWM) models was inferior with the highest false positive rate. The best recognition efficiency was achieved by the combination of SiteGA & diChIPMunk/ChIPMunk models, properly identifying FoxA BSs in up to 90% of loci for both ChIP-Seq datasets.

Conclusions: The experimental verification of transcription factor binding sites models by the study of TF binding to oligonucleotides corresponding to predicted sites increases the reliability of computational methods for TFBS-recognition in ChIP-Seq data analysis. For the ChIP-Seq data interpretation, basic PWMs have inferior TFBS recognition quality compared to the more sophisticated SiteGA and de novo motif discovery methods. A models combination from different principles allowed identification of proper TFBSs [1].

Acknowledgements: This work was supported by budgetary projects VI.60.1.2, VI.58.1.2 and Program 6.6 “Molecular and Cell Biology” from the Presidium of RAS.

References:

1. Levitsky V.G. *et al*., (2014) Application of experimentally verified transcription factor binding sites models for computational analysis of ChIP-Seq data, BMC Genomics, 15(1):80.

DATABASE OF QUANTITATIVE CHARACTERS OF PROCESSES IN EMBRYONIC STEM CELLS

Oshchepkova E.A.*, Omelyanchuk N.A., Akberdin I.R., Ermak T.V., Afonnikov D.A.
Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
e-mail: nzhenia@bionet.nsc.ru
**Corresponding author*

Key words: ESC, embryonic stem cells, database, quantitative characters, gene expression, constants, concentration dynamics

Motivation and Aim: Easy and integral access to quantitative data, such as constants of biochemical reactions, concentrations of mRNAs and proteins in cells and their dynamics are very helpful for planning experiments, discussing and assaying their results and for modelling biological processes *in vivo*, *in vitro* and *in silico*. However, these data are dispersed in thousands of scientific publications and represent poorly structured field of knowledge. Existing databases generally accumulate information on structure of genes, metabolic pathways, enzyme processes; it is also possible to find constants of biochemical reactions, weigh of proteins, gene sequences lengths. Thus, the large part of quantitative characteristics is rarely if ever mentioned in existing databases. Here we describe our approach to create a new database, allowing systematic accumulation of a wide range of quantitative parameters. As an example of this, the database of experimental data on quantitative parameters, characterizing induction of pluripotent stem cells, their self-renewal and differentiation was developed.

Results and Conclusion: The database DynNet, is a manually curated database, which contains information on changes in mRNA and protein concentrations and other parameters of biomolecular processes in ESCs and iPSC. First version of the database contains the data on mRNAs and proteins of human and mouse *POU5F1, NANOG, SOX2* genes: constants of their interactions with other gene products; half-life constants; concentrations in cells and dynamics of concentration changes in different cellular processes (for example, during a differentiation of cells), etc.

DynNet may be used not only as information resource, but also can be applied as a significant part of a computational pipeline for reconstruction of mathematical models, in particular, designed for study different molecular genetic aspects of ESCs ability to maintain their self-renewal and pluripotency.

DEPPDB – A PORTAL FOR ELECTROSTATIC AND OTHER PHYSICAL PROPERTIES OF GENOME DNA

Osypov A.A.*
Laboratory of Cell Genome Functioning, Institute of Cell Biophysics RAS, Pushchino MR, Russia
e-mail: aosypov@gmail.com
**Corresponding author*

Key words: DNA electrostatic potential, physical properties, data integration, genomics

Motivation and Aim: DNA is highly charged and its electrostatic and other physical properties define its shape in the functional space and influence its interactions with different proteins, esp. regulating transcription, in particular RNA-polymerases and TFs. DEPPDB was developed to hold and provide all available information on these properties of genome DNA combined with its sequence and annotation of biological and structural properties of genome elements and whole genomes, organized on a taxonomical basis.

Methods and Algorithms: The electrostatic potential around the double-helical DNA molecule was calculated by the original method [1] using a program package [2,3]. Calculations of other physical properties are based on the di- and trinucleotide content. Different cross-correlation analysis algorithms are applied.

Results: DEPPDB contains all completely sequenced bacterial, viral, mitochondrial, plastids and eukaryotic genomes according to current release of NCBI RefSeq [4]. Data for promoters, regulation sites, binding proteins *etc.* are incorporated from DBs and literature. All data are fully integrated, several tools are provided to support different forms of analysis. Calculation on the fly of the user-provided sequences is available.

DEPPDB can be considered as a portal or collection of databases on the electrostatic and other physical properties of whole genomes and different genome elements in different taxa and organisms: Promoter DB, Regulatory Sites (Transcription Factors, TF) DB, Gene Starts DB, Terminator DB, *etc.* as well as comprehensive analysis toolbox.

Availability: The database [3] is available for academic use via the web interface at http://deppdb.psn.ru or http://electrodna.psn.ru.

Acknowledgements: The authors are grateful to Saveljeva E.G. for technical support and the Institute of Mathematical Problems of Biology of RAS for hosting the Database.

References:

1. R.V. Polozov et al. (1999) Electrostatic potentials of DNA. Comparative analysis of promoter and nonpromoter nucleotide sequences, J. Biomol. Struct. Dyn., 16(6), 1135-43.
2. A.A. Osypov, G.G. Krutinin, S.G. Kamzolova. (2010) DEPPDB – DNA Electrostatic Potential Properties Database. Electrostatic Properties of Genome DNA, J Bioinform Comput Biol., 8(3): 413-25.
3. A.A. Osypov, G.G. Krutinin, S.G. Kamzolova. (2012) DEPPDB – DNA Electrostatic Potential Properties Database. Electrostatic Properties of Genome DNA elements, J Bioinform Comput Biol, 10(2) 1241004.
4. The NCBI handbook [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 2002 Oct. Chapter 18, The Reference Sequence (RefSeq) Project. Available from http://www.ncbi.nlm.nih.gov/books/NBK21091

DNA PHENOTYPE AND BIOPHYSICAL BIOINFORMATICS OF TRANSCRIPTION REGULATION IN PROKARYOTES: THE ROLE OF ELECTROSTATICS AS A NATURAL SELECTION FACTOR

Osypov A.A.*
Laboratory of Cell Genome Functioning, Institute of Cell Biophysics RAS, Pushchino MR, Russia
e-mail: aosypov@gmail.com
**Corresponding author*

Key words: DNA electrostatic potential, physical properties, data integration, genomics

Motivation and Aim: Genome DNA physical properties (PP) define its shape in the functional space and influence its interactions with proteins, esp. for transcription regulation (TR). DNA is highly charged and electrostatics (E.) contributes greatly to the subject. DEPPDB was developed to provide all available information on these properties of genome DNA combined with its sequence and annotation of biological and structural properties of genome elements and whole genomes, organized on a taxonomical basis.

Methods and Algorithms: DEPPDB and its tools [1, 2] were used to carry out the analysis.

Results: E. potential (EP) is distributed non-uniformly along DNA molecule and correlates, though not corresponds exactly, with GC content, strongly depending on the sequence arrangement and its context (flanking regions). Binding frequency of RNA polymerase to DNA along the genome, measured in direct experiment, correlates to the calculated EP.

TR areas have EP and other PP peculiarities. Binding sites of transcription factors of different protein families in different taxa are located in long areas of high EP. Promoters in average have high value and heterogeneity of EP profile. The transcription starting sites of prokaryotic genomes are characterized by extensive (hundreds of bp) zone of high EP and some peculiarities directly around TSS. This is associated with protein binding and formation of PP due to transcription machinery. Specific details of the TSS EP architecture are similar in related taxa. Promoters up-element demonstrates electrostatic nature.

E. effects on genome functioning interact with other PP of DNA, in particular - bending, thermal stability, supercoiling. They may interact with E. in both, formation, and TR.

Conclusion: E. plays important and universal role in TR in prokaryotes, affecting proteins binding probability and positioning accuracy. It may influence horizontal gene transfer, TR systems evolution and contribute to genome regulatory regions high AT content.

PP formation principles affect such fundamental problems as Chargaff's II rule, redundancy of the genetic code, neutrality of synonymous substitutions; and justify the fundamental idea of DNA phenotype, defining the new principle of biophysical bioinformatics.

Availability: DEPPDB is available at http://deppdb.psn.ru or http://electrodna.psn.ru.

Acknowledgements: The authors are grateful to IMPB RAS for hosting the DB.

References:

1. A.A. Osypov, G.G. Krutinin, S.G. Kamzolova. (2010) DEPPDB – DNA Electrostatic Potential Properties Database. Electrostatic Properties of Genome DNA, J Bioinform Comput Biol., 8(3): 413-25.
2. A.A. Osypov, G.G. Krutinin, S.G. Kamzolova. (2012) DEPPDB – DNA Electrostatic Potential Properties Database. Electrostatic Properties of Genome DNA elements, J Bioinform Comput Biol, 10(2) 1241004.

IDENTIFYING OVERREPRESENTED BIOLOGICAL PROCESSES IN COSMONAUTS ON THE FIRST DAY AFTER LONG DURATION SPACE FLIGHTS BY STUDYING URINE PROTEOME

Pastushkova L.H.[1], Kononikhin A.S.[1, 2], Tiys E.S.[3], Obraztsova O.A.[1], Dobrohotov I.V.[1], Kireev K.S.[4], Ivanisenko V.A.[3], Nikolaev E.N.[2], Larina I.M.[1]

[1]Russian Federation State Scientific Center - Institute for Biomedical Problems RAS, Moscow, Russia;
[2]Institute for Energy Problems of Chemical Physics RAS, Moscow, Russia;
[3]Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia;
[4]Gagarin Cosmonauts Training Center, Star City, Moscow region, Russia

Physiology together with information technology approaches will help to understand the etiology and pathogenesis of various diseases, as well as responses and adaptation to internal and environmental factors. There are some physiological effects immediately after long duration space flight during acute phase of readaptation to the Earth's gravity. Their phenomenology is well investigated for more than five decades. However, the formation mechanisms of the majority of the effects are still unknown. We attempted to describe biological processes based on changes in the proteome composition of the cosmonauts' urine.

The urine protein composition samples of 15 Russian cosmonauts (male, aged of 35 up to 51) performed long flight missions and varied from 169 up to 199 days on the International Space Station (ISS) were analyzed. Urine samples were collected before launch and on the first and seventh days after landing, from the second morning fraction (space experiment «Proteome»). As a control group, urine samples of 12 back-up cosmonauts were analyzed. This group is epidemiologically identical to the cosmonaut cohort. Sex, age and protocol of investigation were the same to the prime cosmonauts group. Sample preparation was performed via liquid chromatography-mass-spectrometry. Liquid chromatography-mass-spectrometry was performed on a nano-HPLC Agilent 1100 system (Agilent Technologies Inc., USA) in combination with a LTQ-FT Ultra mass spectrometer (Thermo Electron, Germany). 294 proteins were detected. 34 of these proteins were specific for the first post-flight day. Using BiNGO tool software we found 63 overrepresented processes on the first day after landing ($p < 0.05$ with consideration of Benjamini-Hochberg correction for multiple comparisons). 9 processes were identified on the first day after spaceflight, which were related to renal function and water-electrolytic balance. 3 processes were related with cardiovascular system adaptation to the Earth's gravity, and 9 processes were associated with lipid peroxidation, regulation of reactive oxygen species metabolic process. Thus, it was possible to identify proteins participating in physiological processes which provide adaptation to microgravity and readaptation to the ground conditions.

IDENTIFICATION OF *DROSOPHILA MELANOGASTER* GENES AS BIOSENSORS OF CHEMICAL POLLUTANTS (FORMALDEHYDE, DIOXIN, TOLUENE) AND GAMMA-IRRADIATION IN LOW DOSES

Peregudova D.O.[1]*, Plyusnina E.N.[1, 2], Shaposhnikov M.V.[1, 2], Kudryavtseva A.V.[3], Snezhkina A.V.[3], Moskalev A.A.[1, 2, 4]

[1]*Institute of Biology of Komi SC Ural Branch RAS, Syktyvkar, Russia;*
[2]*Syktyvkar State University, Syktyvkar, Russia;*
[3]*Engelhardt Institute of Molecular Biology of RAS, Moscow, Russia;*
[4]*Moscow Institute of Physics and Technology (State University), Dolgoprudny, Russia*
e-mail: delovoyd3@gmail.com
**Corresponding author*

Key words: gene expression, transcriptome, dioxin, toluene, formaldehyde, low doses of gamma-irradiation, *Drosophila melanogaster*

Motivation and Aim: Organisms are constantly exposed by adverse environmental factors in low doses. Differences in the patterns of gene expression that depend on the influencing factors can be used to biosense each of these pollutants. The aim of this study was to determine exposure-specific genes that can be used as biosensors of chemical pollutants (formaldehyde, dioxin, toluene) and of γ-irradiation in low doses.

Methods and Algorithms: We used a systems approach to analyze the effects of low doses of γ-radiation (20 cGy), formaldehyde vapors (7%), dioxin (0.822 mM), and toluene (50 mM) on the expression of 25,415 transcripts obtained by RNA-seq analysis in adult wild-type (*Canton-S*) *Drosophila* males and females. Thereafter, RT-PCR analysis of selected genes with significantly modified expression was carried out after exposure of γ-radiation (10, 20, 40 cGy), formaldehyde vapors (3.5, 7, 14 %), dioxin (0.016, 0.822, 1.644 mM) and toluene solution (1, 50, 100 mM).

Results: An analysis of genes unique to each treatment yielded a list of genes as a gene expression signature. In the case of radiation exposure, both sexes exhibited an increase expression of transcription factor genes *sugarbabe* and *tramtrack*. The influence of dioxin up-regulated metabolic genes *anachronism*, *CG16727,* and several genes with unknown function. Toluene activated a detoxication gene *Cyp12d1-p*; the transcription factor *Fer3*'s gene; the metabolic genes *CG2065, CG30427,* and *CG34447*; and the genes *Spn28Da* and *Spn3*, which are responsible for reproduction and immunity. An analysis of selected genes identified *tramtrack* as biosensor of γ-radiation.

Conclusion: In this work we obtained gene expression signatures of chemical pollutants (formaldehyde, dioxin, toluene) and γ-irradiation in low doses and common molecular pathways for different kind of stressors. We identified the *tramtrack* gene as biosensor of γ-radiation. However, biosensor genes for chemical pollutants didn't be identified. Therefore, it is required to extend a range of studied genes.

Availability: This study focuses on the mining of biosensors of chemical pollution and gamma-irradiation in low doses.

Acknowledgements: The work was supported by grant of the Presidium of RAS № 12-P-4-1005 "Ecological genetics of lifespan in animal models (*Drosophila melanogaster, Mus musculus*)", RFBR grant № 14-04-01596, Grant of the President of Russian Federation № MD-1090.2014.4 and grant of the Ural Branch of RAS for young scientists № 14-4-NP-103.

THE STATISTICAL ANALYSIS OF LEVEL OF AN EXPRESSION OF A SERIES OF PROTEINS OF A PLANT OF *ARABIDOPSIS THALIANA* IN THE CONDITIONS OF STRESSFUL INFLUENCE

Perfileva A.I.
Federal State Budget Science Institution Siberian Institute of Plant Physiology and Biochemistry SB RAS, Irkutsk, Russia;
National Research Irkutsk State Technical University, Irkutsk, Russia
e-mail: alla.light@mail.ru

One of the assumptions of the world scientific society is that the development of plant resistance to biotic and abiotic stresses is of diverse character (Maimbo et al., 2007; Kilian et al., 2007). Plant cells under abiotic (heat) stress are known to synthesize heat shock proteins protecting the cells for further stress factors (Wang et al., 2004). Under biotic stress plant cells synthesize PR-proteins protecting the cell from consequences of biotic stress. Presumably, provided PR-proteins are synthesized in the cell, HSP synthesis if suppressed. Databases on gene expression in plants under the impact of various stress factors contain numeric values of the expression of genes encoding PR-proteins and HSP.

Bioinformation analysis of Arabidopsis genes expression in response to heat stress carried out with the use of database AtGenExpress (http://jsp.weigelworld.org/expviz/expviz.jsp) on the basis of the results by Kilian et al. (2007), demonstrated that activation of HSP genes under heat stress is, as a rule, accompanied by the suppression of PR-proteins genes expression. The work was aimed at the correlation analysis of the expression of genes encoding HSP and genes encoding PR-proteins in roots, stems and cell cultivar of Arabidopsis (*Arabidopsis thaliana*) under heat stress (38 ^{o}C) with the use of database AtGenExpress. Correlation analysis of the data was carried out with the software Statistica 6.0. The analysis allowed to produce correlation matrices.

The correlation analysis of primary data manifested dependence of the proteins correlation on the object of study. Thus, we observed no interconnection between HSP and PR-proteins synthesis under stress in Arabidopsis stems and roots. Stems and roots are characterized by the presence of diverse tissues with different metabolism intensity. The presupposed initially negative correlation between PR- proteins and HSP synthesis was, through the correlation analysis, identified only in Arabidopsis cell cultivar. In Arabidopsis cell cultivar the cells are totipotent and homogenous, so the studies conducted on this object are more reliable, which is also confirmed by the correlation analysis. Therefore, negative correlation between PR-proteins and HSP synthesis confirms the hypothesis on the presence in the plant of two protective programs with different directions.

References:

1. Kilian J. The AtGenExpress global stress expression data set: protocols, evaluation and model data analysis of UV-B light, drought and cold stress responses / J. Kilian, D. Whitehead, J. Horak, D. Wanke, S. Weinl, O. Batistic, C. D'Angelo, E. Bornberg-Bauer, J. Kudla, K. Harter // Plant J. - 2007. - V. 50, № 2.- P. 347-363.
2. Maimbo M. Induction of a small heat shock protein and its functional roles in Nicotiana plants in the defense response against *Ralstonia solanacearum* / M. Maimbo, K. Ohnishi, Y. Hikichi, H. Yoshioka, A. Kiba // Plant Physiol. - 2007. - V. 145, № 4.- P. 1588-1599.
3. Wang W. Role of plant heat-shock proteins and molecular chaperones in the abiotic stress response / W. Wang, B. Vinocur, O. Shoseyov, A. Altman // Trends Plant Sci. - 2004. - V. 9, № 5.- P. 244-252.

MATHEMATICAL MODELING OF PEPETIDOGLYCAN PRECURSOR BIOSYTHESIS IN THE CYTOPLASM OF *ESCHERICHIA COLI* CELL

Perfilyeva O.A.[1]*, Likhoshvai V.A.[1, 2], Khlebodarova T.M.[1]
[1]*Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia;*
[2]*Novosibirsk State University, Novosibirsk, Russia*
e-mail: perf@bionet.nsc.ru
**Corresponding author*

Key words: peptydogycan biosynthesis, cell growth

Motivation and Aim: Determination of bacterial cell growth strategy is of fundamental interest and plays a key role in solving the problem of coordinating cell growth and replication of the genome [1]. Inability to determine cell growth law with experimental methods [1, 2] addresses the problem to theoretical studies. Mathematical modeling of peptidoglycan synthesis dynamics, which is the main component of the cell wall and defines cell growth, is one of the possibilities to solve the problem. Developing peptidoglycan biosynthesis mathematical model of the cytoplasmic stage in E.coli is the first step in determining the growth strategy.

Methods and Algorithms: The modeling process was performed using chemical kinetic approach, as well as King-Altman method. Genetic algorithms on the basis of pulsed nonequilibrium kinetics data and characteristics of enzyme reactions, such as K_m (Michaelis constant) and V_{max}, were used to estimate the elementary reaction model parameters.

Results: An elementary reaction models database of nucleotide peptidoglycan precursor synthesis catalyzed by enzymes MurA, MurB, MurC, MurD, MurE and MurF was developed. Based on it equilibrium and nonequilibrium peptidoglycan biosynthesis models of cytoplasmic stage, which used the concentrations of the enzymes as additional variable, were constructed. The value of the concentrations was estimated according to the accumulation of peptidoglycan precursor during the cell cycle. The numerical analysis of negative regulation loop influence on activity of enzymes MurA, MurC, MurD, and MurE by intermediate products of nucleotide peptidoglycan precursor synthesis was made.

Conclusion: A model displays a stable kinetic of intermediate peptidoglycan precursor products synthesis in the cytoplasm of the cell and can be used to study the cell growth laws.

Acknowledgments: The work was supported by the Program of Siberian Branch of the Russian Academy of Sciences (Integration project № 80)

References:

1. Likhoshvai V.A., Khlebodarova T.M. (2014) Mathematical modeling of bacterial cell cycle: The problem of coordinating genome replication with cell growth. J. Bioinform. Comput. Biol. V. 12. № 3. 1450009.
2. Cooper S. (2006) Distinguishing between linear and exponential cell growth during the division cycle: Single-cell studies, cell-culture studies, and the object of cell-cycle research. Theor. Biol. Med. Model. V. 3. P. 10.

miRNAs BINDING TO mRNAs OF RICE MYB GENES

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

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

Motivation and Aim: Transcription factors (TF) play a key role in gene expression regulation. 2025 transcription factors of *Oryza sativa L. ssp. indica* are included in 63 families. The family of MYB TF genes plays a key role in the plant growth, development and stress reactions. 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 *O. sativa* MYB genes, which genes are targets for each miRNA family and calculate free energy of miRNA-mRNA interaction.

Methods and Algorithms: The binding sites of osa-miRNA in mRNA of 136 gene of MYB family was revealed using the program MirTarget, which defines the location of the miRNA with mRNA 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 miRNA with mRNA selected by $\Delta G/\Delta G_m$ equal or above to 85%.

Results: At present 661 miRNAs and 136 MYB TF genes are known in the genome of rice. It was found that only 173 miRNAs bind to mRNAs of *O. sativa* MYB genes with value of score ranging from 85% to 96% and only 86 genes are controlled by miRNAs. Osa-miR2102-5p has the largest number of target genes. It bound to mRNAs of 23 genes with value of score from 86 to 91%. Osa-miR5075 bound to mRNAs of 11 genes with value of score equal 86-93%. In the genome of rice some miRNAs bound to mRNAs of one or several genes. Six miRNAs of osa-miR159a-f family bound to mRNAs of OsIBCD015177, OsIBCD018450, OsIBCD021789 genes with value of $\Delta G/\Delta G_m$ equal 87-96%. All ten miRNAs of osa-miR439a-j family bound to mRNAs of only one gene OsIBCD007874 with the same value of $\Delta G/\Delta G_m$ equal 85%. Six miRNAs of osa-miR164a-f family bound to mRNAs of OsIBCD014655, OsIBCD001147, OsIBCD000203 and OsIBCD028679 genes with values of $\Delta G/\Delta G_m$ equal 87-96%. Osa-miR169a-e bound to mRNAs of five genes (OsIBCD015070, OsIBCD022880, OsIBCD003079, OsIBCD032138 и OsIBCD008009) with score equal 85-87%. Osa-miR159 bound to mRNAs of three genes (OsIBCD015177, OsIBCD018450, OsIBCD021789) with value of score ranging 87-96%. Osa-miR1846d-3p bound to mRNAs of OsIBCD018922, OsIBCD023878, OsIBCD026910 and OsIBCD035359 genes. All rice miRNA binding sites are located in the protein-coding part of MYB TF family mRNAs.

Conclusion: As a result of study of *O. sativa* MYB 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 MYB family genes are targets for miRNA; 3) nucleotides of miRNA binding sites encode oligopeptides of MYB proteins in different reading frames. Results obtained show that osa-miRNAs can regulate expression the most of MYB genes and influence the plant productivity and resistance.

THE MAMMALIAN CIRCADIAN CLOCK: COMPUTER ANALYSIS OF GENE NETWORK

Podkolodnaya O.A.[1]*, Podkolodnaya N.N.[1], Podkolodnyy N.L.[1, 2]
[1]*Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia;*
[2]*Institute of Computational Mathematics and Mathematical Geophysics SB RAS, Novosibirsk, Russia*
e-mail: opodkol@bionet.nsc.ru
**Corresponding author*

Key words: circadian rhythm, graph analysis, regulatory circuits

Motivation and Aim: Circadian clocks are universal adaptive mechanism in eukaryotes that oscillates with a period of 24 hours. They provide timing in living organisms of processes at all levels from the molecular genetic to physiological and behavioral. The cyclical nature of the gene expression in different mammalian tissues is shown for about 10 thousand genes. That is why the analysis of expression of these genes is not possible without the analysis of influence of core circadian rhythm genes which provide the circadian oscillation.

Methods and Algorithms: To search for strongly connected component and regulatory circuits in gene network Tarjan algorithm was used. Search of putative target genes of transcription factors was carried out by the method of computer prediction of binding sites of these factors, using the weight matrix. Functional annotation of genes was performed using a system DAVID.

Results: Gene network (GRN) "Circadian Rhythm", was reconstructed using the gene networks editor GeneNet on the basis of information from 279 scientific publications. This gene network contains a description of 77 genes, 153 protein, 19 simple compounds, 526 reactions and regulatory events. The search and analysis of regulatory circuits in GRN "Circadian Rhythm" was carried out. The 2 500 000 cycles have been found. Such a large number of regulatory circuits in this network may reflect the importance of the stable functioning of a circadian oscillator which is provided by multi-level control. Notes that more than 20% of the target genes of the main transcription factor circadian oscillator Clock/Arntl are also transcription factors in GRN "Circadian Rhythm". This can provide an oscillating expression pattern for large set of genes due cassette type of regulation of their transcription with the relevant transcription factors, which are under the direct control of circadian oscillator. Sets of genes that are direct potential targets of each of these transcription factors have been compiled on the basis of recognition of their binding sites in the 5' sequences of mouse genes (NCBIM37). Functional annotation of these sets of genes revealed that a wide range of biological functions of the cell may be under the direct control of the cellular circadian oscillator at the transcriptional level.

Conclusion: Regulation of gene expression at the transcriptional level is potentially a powerful mechanism for the circadian control cell function. This should be considered when interpreting the expression data.

AN EMPIRICAL EQUILIBRIUM EQUATION OF A GENE RESPONSE TO AUXIN IN PLANTS ALLOWS TO PREDICT QUANTITATIVELY THE AUXIN RESPONSE UPON THE GENE PROMOTER SEQUENCE

Ponomarenko P.M.[1], Ponomarenko M.P.[2] *
[1]*Children's Hospital Los Angeles, Los Angeles, CA 90027, USA;*
[2]*Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia*
e-mail: pon@bionet.nsc.ru
**Corresponding author*

Key words: plant gene, auxin response element, empirical equilibrium equation, forecast

Motivation and Aim: Auxin is a hormone that regulates plant development via the plant gene activation and/or repression. A prediction of gene response to auxin in plant upon the gene promoter sequence is important problem in bioinformatics.

Methods and Algorithms: First, we analyzed the all six available experimental datasets on quantitative rates of the individual impact of the experimentally proven auxin response elements (AuxREs) into the auxin response of plant genes taken from [1-3]. Upon these datasets taken all together, we calculated both astandard position-frequency matrix and consensus at Boltzmann's approximation. Using our earlier developed system Activity [4], upon each dataset taken separately from the others, we next found the significant correlations between the quantitative rates of each AuxRE impact into auxin response and the contextual DNA features near this AuxRE. Finilly, we comprehensively cross-validation tested in significance each contextual feature found within each dataset treated.

Results: The AuxRE consensus found was equal to the canonical motif TGTCNC known from the *in vitro* experiment [3] with transcription factor ARF1 and synthetic oligo DNA. For each AuxRE impacts into auxin response, we found significant correlations to the number of matches to our consensus, to Homology score [5] of our position-frequency matrix, to the averaged values of twist, slide, tip, inclination, minor groove size of DNA helix, and to abundances of triplets TSD, HYR, YHD, VVT, SNW near this AuxRE. That was the only one reason why we generalized the all twelve contextual features found near AuxRE into an empirical equilibrium equation describing four steps of auxin response in plant. Fist, at the absence of auxin, ARF•Aux/IAA heterodimer jumps from site to site along DNA and checks up its key-lock fitness by its unspecific ARF/DNA-affinity. Then, these jumps stop at AuxRE/ARF•Aux/IAA. Next, Aux/IAA represses the basal transcription of the gene containing this AuxRE. Finally, as soon as auxin appears, Aux/IAA dissociates and degrades, whereas AuxRE/ARF induces the gene transcription.

Conclusion: We verified our empirical equilibrium equation by 70 arabidopsis auxin-responsive genes with unknown AuxREs. It resulted by both linear and rank significant correlations between the auxin response rates predicted and those measured elsewhere.

Acknowledgment: We thank RFBR-12-04-33112-mol-a-ved & Budget Project VI.61.1.2.

References:

1. V. Mironova et al. (2013) How multiple auxin responsive elements interact in plant promoters: evidences from a reverse problem solution. JBCB. 11:1340011.
2. T. Ulmasov et al. (1995) Composite structure of auxin response elements. Plant Cell. 7: 1611-1623.
3. T. Ulmasov et al. (1997) ARF1, a transcription factor that binds to auxin response elements. Science. 276: 1865-1868.
4. M. Ponomarenko et al. (1997) Generating programs for predicting the activity of functional sites. J. Comput. Biol. 4: 83-90.
5. D. Hawley and W. McClure (1983) Compilation and analysis of Escherichia coli promoter DNA sequences. Nucleic Acids Res. 11: 2237-2255.

SCAN ELEMENTS IN THE NON-CODING DNA

Ponomareva N.S.*, Romanov D.E., Pshenichny E.A., Shkurat T.P.
Southern Federal University, Rostov-on-Don, Russia
e-mail: nsponomareva@sfedu.ru

Key words: Non-coding DNA, DNA Motif, Scan Motif

Motivation and Aim: Non-coding DNA was begun research in the last 5 years. Existing software is targeted on research within the protein-coding sequences. The problem of automatic data extraction, conversion formats, search motives and the resulting formation of databases are substantial in the study of non-coding DNA. Study noncoding DNA regions: introns and areas between genes to determine their structure and function in processes regulation.

Methods and Algorithms: Program software modules IFetch and Charlie to automatically extract through the identification number of the gene sequences of non-coding DNA data bank NCBI http://www.ncbi.nlm.nih.gov/gene and save as FASTA-file, with user-friendly file names were designed.

The IFetch module automatically saves areas before and after the gene name and generates the file name of the gene sequence and specify the location, relative to the gene.

The Charlie module extracts intron sequences and saves every sequence generates the file name of name of the gene and number of the gene intron.

Mscanner module for automatically search the different motives of created an integrated database of functional elements by using GLAM2 and stores the results in a database was designed. Name of new database specified by the user. SQL queries extract data for further analysis. This paper presents similarity index, showing the level of similarity with the motif of the base found in the non-coding sequences.

Results: Non-coding sequences before and after Titin gene functional elements mir-669n, mir-2325c, mir-34, mir-2359, mir-5193 were detected. For instance, mir-519 comprises a nucleotide sequence aggaggagatggagtagggtca, and non-coding sequences before Titin positions 10367 - 10386 is the sequence aggagCag..ggagtagggtca capital letter denotes the substitution, insertion is indicated by point, the index of similarity is 0.86.

Conclusion: software modules that automatically retrieve through the identification number on the gene sequence of noncoding DNA protein in and around the gene search of different motifs to create an integrated database of functional elements by using GLAM2 and save the results in a database for further analysis were developed. Discovered motifs extend the genetic information of the structures contained in the DNA sequences. Identify patterns in the organization and functioning of the genome, it is necessary to conduct a search in the intergenic spaces around and within the same genes in different organisms and different groups of genes, which proved the dependence of functioning is main objective.

*Availability***:** software modules are publicly available on the cloud service BioSFedU http://biosfedu.no-ip.org/progs/, access to the cloud service are available from the authors upon the request.

ANNOTATIONS OF SNPS IN PROMOTERS OF HUMAN ONCOGENES VEGFA, EGFR, ERBB2, IGF1R, VEGFR1(2) AND HGFR IN TERMS OF POTENTIAL RESISTANCES TO MONOCLONAL ANTIBODY DRUGS

Popik O.V.[1], Arkova O.V.[1], Rasskazov D.A.[1], Ponomarenko P.M.[2], Titov I.I.[1], Arshinova T.V.[1], Saik O.V.[1], Savinkova L.K.[1], Ponomarenko M.P.[1]*, Kolchanov N.A.[1, 3]
[1]*Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia;*
[2]*Children's Hospital Los Angeles, Los Angeles, CA 90027, USA;*
[3]*Novosibirsk State University, Novosibirsk, Russia*
**Corresponding author e-mail: pon@bionet.nsc.ru*

Keywords: SNP, annotation, TATA-box, TBP, oncogene, mAb-therapy, drug resistance

Motivation and Aim: A search for single nucleotide polymorphisms (SNP) in oncogenes that may be associated with drug resistance in antitumor monoclonal antibody (mAb-) therapy is of high importance because it is necessary to adjust the treatment in this case.

Methods and Algorithms: From dbSNP [1] we took 128 SNPs in-between -20 bp and -70 bp relative to each of the all transcription start site noted by hg19 (the region where all known TATA boxes reside) of human oncogenes VEGFA, EGFR, ERBB2, IGF1R, VEGFR1, VEGFR2, HGFR which are gene-targets of monoclonal antibody drugs [2]. We examined each of them by our empirical equation of TBP binding to TATA box [3] that was verified at both equilibrium [4] and non-equilibrium [5] conditions *in vitro* by significant ($r=0.82$, $p<10^{-3}$) correlations between the predicted and measured K_D-values of TBP/TATA-complex. The SNP-caused decrease in K_D-value related to the gene norm may mean a resistance to the monoclonal antibody drugs in cancer mAb-therapy [2].

Results: We found ten SNPs that significantly decrease K_D-values of the TBP/promoter-complex of human oncogenes VEGFA, EGFR, ERBB2, IGF1R, VEGFR1, VEGFR2, and HGFR. These SNPs may cause drug resistance in cancer mAb-therapy [2]. As an example, SNP -25g → A of the non-canonical TATA-box of human gene VEGFA, cctcccccTT(g → A)GGATcccgca, rs370995111, may up to 4-fold decrease K_D-value, 195 nM → 53 nM, and, thus, may cause an excess of vascular endothelial growth factor encoded by this gene. It may be a reason for a potential resistance to bevacizumab [2]. Also, we found twelve SNPs that significantly increase K_D-values of TBP/promoter-complexes of these oncogenes. Using them may lead to reducing a dose of a proper monoclonal antibody drug and, thus, adding proportionally a dose of another drug to improve mAb-therapy. We annotated 106 SNPs remaining neutral in the sense described above by the mark "an insignificant change in TBP affinity to promoters of the gene".

Conclusion: We first annotated SNPs of oncogenes in a resistance to mAb-therapy drags.

Acknowledgment: We thank RFBR-14-04-00485, Budget Projects VI.58.1.2 & VI.61.1.2.

References:

1. M. Bhagwat (2010) Searching NCBI's dbSNP database. Curr. Protoc. Bioinf. 1, 1.19.
2. O. Polanovski et al. (2012) ERBB oncogene proteins as targets for monoclonal antibodies. Biochemistry (Mosc). 77: 227-245.
3. P. Ponomarenko et al. (2008) A step-by-step model of TBP/TATA box binding allows predicting human hereditary diseases by single nucleotide polymorphism. Dokl Biochem Biophys. 419: 88-92.
4. L.K. Savinkova et al. (2013) An experimental verification of the predicted effects of promoter TATA-box polymorphisms associated with human diseases on interactions between the TATA boxes and TATA-binding protein. PLoS ONE. 8: e54626.
5. I. Drachkova et al. (2014) The mechanism by which TATA-box polymorphisms associated with human hereditary diseases influence interactions with the TATA-binding protein. Hum. Mutat. 35: doi: 10.1002/humu.

EVALUATION OF PATHWAYS' EFFICIENCY BASED ON DATA ON PPI AND DISTRIBUTION OF PROTEINS OVER CELLULAR LOCALIZATIONS

Popik O.V.*[1], Hofestaedt R.[2], Ivanisenko V.A.[1]
[1]*Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia;*
[2]*University of Bielefeld, Germany*
e-mail: popik-olga@bionet.nsc.ru
**Corresponding author*

Key words: gene network, integration

Motivation and Aim: Intracellular protein localization plays important role in cell functioning. Biological processes occurring in the cell also can be distributed over the intracellular compartments. It can be suggested that the efficiency of pathway's functioning will be depended on which compartments the proteins involved are located. In some compartments biological processes might proceed with high rate and in some rate is slow. The rate of the realization of whole pathway will be summarized from intra-compartmental and inter-compartmental rates. We suggested that the frequency of protein-protein intra-compartmental and inter-compartmental interactions is the kind of estimation of pathway rate. There are many databases containing data about PPI and proteins intracellular localizations. ANDSystem [1] integrates protein intracellular localization data and information on PPI, extracted from various databases such as UniProt, IntAct, KEGG, BIND, etc., and scientific publications.

Methods and Algorithms: By using data from ANDSystem we constructed three matrixes of frequency of PPI for all pairs of cellular compartments – for human, for yeasts and for Escherichia coli. The total numbers of localization were 20 for human, 15 for yeasts and 10 for E.coli. The diagonal elements of each matrix show the frequency of intra compartments PPI, out diagonal elements show the frequency of PPI between different compartments. Analysis of the matrixes revealed that the frequency of intra-compartments PPI occurs more often in comparison with inter-compartments PPI.

Results: The matrixes for evaluation the efficiency of pathways and gene networks were constructed.

The human KEGG pathways were evaluated using the matrixes.

Conclusion: The method to evaluate pathways' efficiency using data on PPI and was suggested.

References:

1. P.S. Demenkov, T.V. Ivanisenko, N.A. Kolchanov, V.A. Ivanisenko. ANDVisio: a new tool for graphic visualization and analysis of literature mined associative gene networks in the ANDSystem. In Silico Biol. 2011-2012; 11(3-4):149-61. doi: 10.3233/ISB-2012-0449.

MATHEMATICAL MODELING OF THE INTERACTIONS BETWEEN MOLECULAR GENETIC SYSTEMS BASED ON THE DATA ABOUT PERTURBATIONS OF THE SYSTEMS ELEMENTS

Popik O.V.*, Kolchanov N.A., Ivanisenko V.A.
Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
e-mail: popik-olga@bionet.nsc.ru
**Corresponding author*

Key words: gene network, integration

Motivation and Aim: Any problem of studying complex molecular genetic processes requires solving the problem of integration of simple subsystems. To date there are a lot of mathematical models describing biological processes in normal and pathological conditions of the body. Typically, such models are based on information about the molecular and genetic networks that are highly specialized subsystems at the level of global gene regulatory network of an organism or cell. Common approach is to the integrate models by the intersection of gene networks' elements. Such approach severely limits its application to the processes indirectly influence each other and to the understudied biological processes.

Now a huge amount of available experimental data on the response of biological systems to various perturbations is available. Application of mathematical control theory methods using data on the response of biological systems to various perturbations can solve the problem without explicit reconstruction of indirect links between the simulated biological subsystems.

Methods and Algorithms: Based on modeling of gene networks by systems of ordinary differential equations the method of integrating the two models by control function was developed.

The properties of the method suggested were studied on generated random gene networks on the computer. Networks were divided on sub networks, the function was constructed using perturbations simulated and its properties were analyzed.

Results: The functional was successfully constructed for all generated network. It is shown that construction of the functional requires data about perturbation not less than 2 elements of the network and in the majority of cases such data are enough. Also in some cases it is possible to reconstruct interconnections between sub netwroks.

Conclusion: The study on the generated random networks showed that the construction of the control functional is possible in the absence of information about the structure of a gene network describing the indirect relationship between the two subsystems. On the basis of random gene networks analysis regularities between structural features of the GN and data required to construct the control function were found.

REFERENCE ASSISTED CHROMOSOME ASSEMBLY OF 30 *SACCHAROMYCES CEREVISIAE* STRAINS FROM SACCHAROMYCES GENOME DATABASE

Radchenko E.A.*, Tamazian G.S., Dobrynin P.V.
Saint-Petersburg State University, Saint-Petersburg, Russia
e-mail: elle.radchenko@gmail.com
**Corresponding author*

Key words: *Saccharomyces cerevisiae*, yeast, SGD, genome assembly

Motivation and Aim: Saccharomyces cerevisiae is a hallmark model organism in different fields of modern biology, from genetics to proteomics. Strain S288C was first sequenced eukaryotic genome [1]. There is one high quality chromosome assembly available for public usage and Saccharomyces Genome Database [2] (SGD, www.yeastgenome.org) also contains additional 30 strains assembled to contigs. Anchoring contigs to chromosomes and producing a chromosome assembly is important for addressing questions about chromosome rearrangements and signatures of selection.

Methods and Algorithms: Thirty assemblies of *S. cerevisiae* strains were downloaded from SGD. Sequence coverage for samples were from 2X to almost 400X, the contig length from 683 bp to 376 kbp with N50 from 1 kbp to 800 kbp. As a reference genome sequence we used S288C whole genome chromosome assembly with its mitochondrial genome. We used the in-house software tool to get chromosome assemblies of the considered strains. The tool uses lastz [3] or blastn [4] alignments between fragments of a genome to be assembled and chromosome sequences of a reference genome. Based on the alignments, the fragments are mapped to reference genome chromosomes and thus the chromosome assembly is derived.

Results: We produced 30 chromosome assemblies based on S288C reference genome. For all samples chromosomes were assembled and eight of them with a mitochondrial genome. Quality of each assembly was checked with QUAST[5] and CEGMA[6]. There was no surprise that correlation between high quality scores in QUAST, CEGMA and high sequence coverage were observed. We also used custom python script to transfer genome annotation to newly assembled chromosomes.

Availability: All produced assemblies and their annotations are available at Dobzhansky Center Hub for UCSC Genome Browser (http://public.dobzhanskycenter.ru/Hub/hub.txt).

References:

1. A. Goffeau et al. (1996) *Science*, 274: 546-567.
2. J.M. Cherry et al. (2012) *Nucleic acids research* 40.D1: D700-D705.
3. R.S. Harris. (2007) *ProQuest*.
4. S.F. Altschul et al. (1990) *Journal of molecular biology*, 215.3: 403-410.
5. A. Gurevich et al. (2013) *Bioinformatics*, 29.8: 1072-1075.
6. G. Parra et al. (2007) *Bioinformatics* 23.9: 1061-1067.

IDENTIFICATION OF POTENTIAL INHIBITOR OF PROTEIN KINASE D1 (PKD1) AND 2 (PKD2)

Raevsky A.V., Samofalova D.A.*, Karpov P.A., Blume Y.B.
Institute of Food Biotechnology and Genomics, Natl. Academy of Sci. of Ukraine, Kyiv, Ukraine
e-mail: samofalova.dariya@gmail.com; rayevsky85@gmail.com
**Corresponding author*

Key words: inhibitor of PKD, ligand - protein interactions, molecular docking and dynamics

Motivation and Aim: The protein kinase D (PKD) family of serine/threonine protein kinases which play a role in metastasis, immune responses, apoptosis and cell proliferation. The PKD family contains 3 members that are homologous in structure and function, namely, PKD1, PKD2, and PKD3. However, all members of the family share distinct structural homology. The human PKD1 is the largest member, with 912 amino acids and a molecular mass of 115 kDa. The other 2 members are PKD2 with 878 amino acids (molecular mass 105 kDa) and PKD3 (previously called PKCν/PKCnu) with 890 amino acids (molecular mass 110 kDa). All PKD members have a common modular structure consisting of an *N*-terminal region with regulatory domains and a C-terminal region represented with a kinase domain.

There are some interesting features of this kinases, which had to be taken into account during 3D-modelling. First, the crystallographic structures currently aren't available and there are only three known selective inhibitors at the time: CID 755673 (IC_{50} values are 0.182 and 0.280 μM at PKD1 and PKD2) and kb NB 142-70 (IC_{50} values are 28.3 and 58.7 nM for PKD1 and PKD2) - allosteric inhibitors which binding site is located within a unstructured region of the C-end, and ATP-competitive CID 2011756 (IC_{50} values are 0.6, 0.7 and 3.2 μM for PKD2, PKD3 and PKD1 respectively).

Methods and Algorithms: To deal with the task of searching for potential inhibitors of PKD1 and PKD2 a procedure of structure modeling was carried out. Based on PDB-BLAST search 5 template structures of known kinases (PDBID: 1ql6, 2bdw, 3mfr, 2jam, 2y7j) were selected for modeling. These templates were used as inputs for I-TASSER server. Only two of 10 structures (580-840 a.a. of catalytic domain) were selected based on ANOLEA ProCHECK values.

Results: To obtain the most favorable PKD1 and PKD2 conformations, GROMACS package for molecular dynamics and amber99sb force field were used. Free MD simulation continued for 20 ns on 32 core node of IFBG cluster. As a positive control several template kinases (PDBID: 1ql6, 3mfr) were also processed to determine if duration of MD is realistic enough. All four MD trajectories were analyzed (rmsd, rmsf) and clustered. For each model up to 20 cluster groups were obtained. Each group was automatically created basing on processing of rmsd matrices and mean distribution of conformations during MD.

Flexible docking ATP and CID 2011756 in ATP-binding site was carried out with GOLD CCDC. It was necessary to test our models and select the most suitable conformations. To increase accuracy of docking procedure several rotamer libraries were assigned to key amino acids. Best poses were selected basing on ASP and GoldScore. After this stage a Dock 4.2 program was used to generate electrostatic maps and spherical volumes. A subset of compounds was prepared with OpenBabel and screened against our dock-models. This screening showed that only 6 compounds of 3000 would be useful for subsequent ligand design. However, all of them are not absolutely selective and the idea of application of other tools (QSAR, phrmacophore) to focus on allosteric site, located in the part of structural disorder, is still actual.

Acknowledgments: The research was supported by "The state scientific and technical program to develop advanced technologies for domestic pharmaceuticals for human health and meet the needs of veterinary medicine in 2011-2015", grant #37/13A NAS of Ukraine.

DISCOVERY OF THE ROLE OF HORIZONTAL GENE EXCHANGE IN EVOLUTION OF PATHOGENIC MYCOBACTERIA

Reva O.[1]*, Korotetskiy I.[2], Ilin A.[2]
[1]*Bioinformatics and Computational Biology Unit, Biochemistry Department, University of Pretoria, Lynnwood Rd., Hillcrest, Pretoria 0002, South Africa;*
[2]*Scientific Center for Anti-infectious Drugs, Almaty, Kazakhstan*
e-mail: oleg.reva@up.ac.za; scaid@scaid.kz
**Corresponding author*

Key words: *Mycobacterium*, genome linguistics, horizontal gene exchange

Motivation and Aim: Important genes may flux among bacteria by horizontal transfer and trigger substantial evolutionary changes. *Mycobacterium tuberculosis* is known to be resistant to horizontal gene transfer. However, acquisition of pathogenicity genomic islands might play role in the evolution of virulence traits of this human super-pathogen.

Methods and Algorithms: Genome lingusitics methods were applied in *Mycobacterium tuberculosis* and several other organisms to study distributions and relationships of their genomic islands (GIs). Complete genome sequences of organisms used in the study were downloaded from NCBI FTP server. Their GIs were identified by SeqWord Genomic Island Sniffer (SWGIS), a tool that detects inserts of foreign DNA in bacterial genomes by comparison of biased distribution of tetranucleotides in the core genome and genomic loci. The identified GIs were clustered by their compositional similarity of oligonucleotide usage patterns. These were further analysed by pairwise blast2seq and blastn/blastp methods where the GIs encoded proteins were searched against local DNA and protein databases of all known bacteria, plasmids and phages. The latter was performed to check if the GIs that cluster together share blocks of genes that are in the same order and to also deduce the types of genes that are highly likely transferred within a genus or species borders.

Results: Novel GI identification approaches helped to identify several GIs in Mycobacteria including virulent and environmental species. Furthermore, six GI identified in *M. tuberculosis*, *M. bovis* and *M. marinum* showed to comprise elements which share similarity in both DNA sequence identity and composition with GIs distributed among alpha-Proteobacteria particularly those of *Rhizobium* and *Agrobacterium*. The identified GIs are distinctive from prophage inserts common for environmental Mycobacteria. Some of the genes of these GIs seem to have been transferred from *Brucella* to *Rhizobium* and *Agrobacterium*, and in a later stage from the latter to *Mycobacterium*.

Conclusion: Acquisition of genetic materials from intracellular parasitic and symbiotic species of alpha-Proteobacteria by an ancestral strain of *Mycobacterium* may be an event that had triggered the evolution of former saprophytic organisms towards the parasitic lifestyle. Practical applicability of newly developed computational tools was demonstrated.

Availability: SWGIS is available for download from www.bi.up.ac.za/SeqWord/sniffer. Identified GIs were stored in Pre_GI database available at http://pregi.bi.up.ac.za/index.php.

Acknowledgments: This work was supported by the research grant #86941 provided by the National Research Foundation (NRF) of South Africa.

EXPERIMENTAL DATA FOR TESTING THE ADEQUACY OF EXISTING MATHEMATICAL MODELS

Ri M.T.[1]*, Wölfl S.[2], Zakhartsev M.[2]
[1]*LLC «AcademGene», Novosibirsk, Russia;*
[2]*Institut für Pharmazie und Molekulare Biotechnologie, Universität Heidelberg, Germany*
e-mail: maxim.ree@googlemail.com
**Corresponding author*

Key words: nucleotides; gene network, mathematical modelling, model verification; *Escherichia coli*, HPLC

Motivation and Aim: Experimental validation of mathematical models is one of the steps in model development in systems biology. Initially we have developed the dynamic model of nucleotide biosynthesis in *E.coli* [1, 2] where stationary intracellular concentrations of metabolites were used for the parameter estimation. The metabolite concentrations were either accepted from different publications for different strains or were estimated. Therefore, our aim was to measure the quasi-steady state intracellular concentrations of the key metabolites in nucleotide biosynthesis for wild type *E.coli* K12 and different mutants (Δ*purR*, Δ*nrdR, rph*-1 frame shift mutation) and carry out the verification of the model [1].

Methods: Bacterial strains and cultivation. Four strains of *E.coli* K12 were cultivated in aerobic batch (wild type, W3110, JW1650-1, JW0403-2) at 37°C on minimal medium supplemented with glucose as a sole carbon source. The biomass was measured by determining the dry mass of the samples and used for values normalization and monitoring of the growth kinetics. *Sampling*. The nucleotide samples were collected in exponential growth phase. The broth was sampled through a heat exchanging coil at 97°C, using fast sampling/quenching techniques. Intracellular metabolites were extracted from the samples, dried out and re-dissolved in buffer A for HPLC analysis. *HPLC*. The ion-pair reversed phase gradient HPLC analysis was performed on the Agilent 1200 Series HPLC system with Supelcosil™ LC-18-T column (buffer A: 100 mM phosphate buffer pH 6.0; 4 mM tetrabutylammonium disulfate (TBAS); buffer B: 30% methanol in 100 mM phosphate buffer pH 7.2 and 4 mM TBAS; gradient of B: 1.5% min^{-1}; flow rate: 1 mL/min; T: 25°C; detection: 260 and 340 nm). The species identification was based on the prior calibration with the purified metabolites (Sigma) and the metabolite contents were expressed in μmol per gram of dry weight of biomass [μmol/gDW].

Results: We have measured quasi-steady state intracellular concentrations of purine and pyrimidine nucleotides (AMP, ADP, ATP, GMP, GDP, GTP, CMP, CDP, CTP, UMP, UDP, UTP, NAD^+ $NADP^+$ and cAMP) by HPLC in four *E.coli* strains (wild type, W3110, JW1650-1, JW0403-2). The experimentally measured steady state concentrations of nucleotides for the wild type strain were used for the re-parameterization of the model [1]. Regulatory protein *PurR* is a key-regulator in the mathematical model [1], therefore the *in silico* mutants *ΔpurR* was estimated using the new parameters. The predicted nucleotide steady state concentrations for the mutant have matched to the experimental ones. According to our knowledge, the experimentally measured UMP concentration for *E.coli* has not been reported yet in the literature. Nevertheless, the model [2] has theoretically predicted it and the experimentally measured value for the wild type strain has matched to the model one.

References:

1. Ri M.T. (2012) Mathematical Modeling and Analysis of *Escherichia coli* Gene Network, Uchenye zapiski the Kazan University. Series of Natural Sciences, 154:247-262.
2. Ree T.M., Khlebodarova T.M., Likhoshvai V.A. (2010) Mathematical modeling of nucleotide biosynthesis in *Escherichia coli* // Seventh international conference on bioinformatics of genome regulation and structure (BGRS/SB_2010), Novosibirsk, Russia, p. 243.

THE MATHEMATICAL MODEL OF Rob, MarR, MarA REGULATORY CIRCUIT OF *ESCHERICHIA COLI* GENE NETWORK

Ri M.T.[1]*, Saik O.V.[2], Khayrulin S.S.[3]
[1]*LLC «AcademGene», Novosibirsk, Russia;*
[2]*Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia;*
[3]*A.P. Ershov Institute of Informatics Systems SB RAS, Novosibirsk, Russia*
e-mail: maxim.ree@googlemail.com
**Corresponding author*

Key words: gene network, mathematical modeling, stochastic systems, oscillations, delayed argument, *Escherichia coli*, OpenCL, parallel computing

Motivation and Aim: Development of an *in silico* cell as a resource for modeling and analysis of intracellular processes is a topical problem of system biology and bioinformatics. Within this direction, it is necessary to develop mathematical models of cell genetic regulation, in particular, a model of regulatory circuit of Rob, MarR and MarA proteins. The aim of this work was to analyze mechanisms of the direct relationships and feedback loops with delay in gene regulation of *E.coli* cell using mathematical modeling.

Methods and Algorithms: Deterministic (D) and stochastic (S) models were developed. D model was constructed as a system of ordinary differential equations describing the rate of protein concentration changes. For the S model without delay Gillespie algorithm was used, the S model with time delay was based on the generalized Gillespie algorithm in the case of non-Markov systems [1]. To implement the parallelized algorithm OpenCL platform was used. The algorithm was implemented in Python and OpenCL C languages and performed on Intel i7 and GPU Tesla C1060 processors.

Results: Results of modeling show that the system may have a steady-state solution and undamped oscillations. Both modes are stable with respect to the intrinsic noise and periodic extrinsic force. Numerical analysis of the model shows that the period of the limit cycle depends on delay argument linearly while the amplitude - nonlinearly. It was revealed that the duplication of the *marR* gene leads to solutions of the complex nature. Parallel computing allowed to reduce the calculation time of the model with no delay by 12 fold and the model with delay by 640 fold [2]. Based on these results we propose a hypothesis about the mechanism of the separation of *E.coli* culture on 2 groups: the one surviving under the influence of a wide range of drugs and the nonsurviving group.

Conclusion: Deterministic and stochastic models of Rob, MarR and MarA regulatory circuit of *E.coli* cell were developed and analyzed.

Availability: Mathematical model, the program code and the results of data analysis are available at https://dl.dropboxusercontent.com/u/52461630/BGRS_2014.zip.

References:

1. D. Bratsun, D. Volfson, L. S. Tsimring, J. Hasty. (2005) Delay-induced stochastic oscillations in gene regulation, *Proc. Natl. Acad. Sci. USA*, **102**: 14593–14598.

2. M.T. Ri, S.S. Khairulin, O.V. Saik. (2013) The mathematical model of Rob, MarA, MarR regulatory circuit of *Escherichia coli* gene network*, Uchenye Zapiski Kazanskogo Universiteta*. Series of Natural Sciences, **155**: 231-249.

FROM 1D INFORMATION TO 3D GENOME STRUCTURE AND FUNCTION

Ruan Y.
Genome Institute of Singapore, 60 Biopolis Street, Singapore 138672
Jackson Laboratory, Genomics Medicine, 400 Farmington Ave, Farmington, CT06030, USA
e-mail: Yijun.Ruan@jax.org

High throughput DNA sequencing technology and the reference genome sequences of human and model organisms signify a significant revolution in modern biology. The comprehensive linear one-dimensional (1D) genome information dramatically changed our views and ways of conducting biological researches. Such momentum of the first waves of genomics leaded to the second waves of genomics prompted by the recent next generation sequencing technologies for cheaper, faster and more sequence data production. Large numbers of individual genome sequencing and in-depth genome interrogation of regulatory elements are rapidly expanding our knowledge in two-dimensions (2D) of genomic information. However, soon we realize that the massive 1D and 2D genomic knowledge do not immediately provide mechanistic answers to many fundamental genetic questions. Expectations are that the mystery of genome functions lies in the three-dimensional (3D) architecture of genome organization. The exciting new challenge is how to study and present the 3D genomes.

It has been viewed that distant DNA elements functionally interact with one another via protein factors. However, the details of this view are largely unchartered. To explore in this direction, we developed the Chromatin Interaction Analysis using Paired-End-Tag sequencing (ChIA-PET) strategy for *de novo* detection of genome wide long-range chromatin interactions, and demonstrated this approach through comprehensive mapping of chromatin interactions mediated by transcription factor estrogen receptor α (ERα) in the human genome [1]. To further map chromatin interactions involved in general chromatin architecture and all transcription regulation, we applied the ChIA-PET analysis to CCCTC-binding factor (CTCF), a chromatin protein associated with chromatin conformation, and RNA polymerase II (RNAPII) that involves in transcription of all genes. Such analyses in human and mouse cells revealed that the general landscape of chromosomal organization is largely framed by CTCF into abundant chromatin looping architectures, which provides the topological framework for transcription coordination and regulation [2, 3]. Overall, our studies paved the way towards presenting the 3D topographic maps of the human genome [4]. The same approaches and concepts can be readily applied for studying the plant genomes to yield new insights of genomic functional controls of complicated agronomic traits.

References:

1. Fullwood, M. et al., (2009). An oestrogen-receptor-α-bound human chromatin interactome. Nature 462: 58-64.
2. Handoko, L. (2011). Functional Genome Architecture Revealed by CTCF-Associated Chromatin Interactome in Pluripotent Cells. Nature Genetics 43: 630–638.
3. Li, G. et al., (2012). Extensive Extensive Promoter-centered Chromatin Interactions Provide a Topological Basis for Transcription Regulation. Cell 148, 84-98.
4. Ruan, Y. (2011). Presenting the Human Genome: NOW IN 3D! Science 331: 1025-1026.

RESTRICTION SITES AVOIDANCE IS TRACE OF LOST RESTRICTION MODIFICATION SYSTEMS

Rusinov I.S.,[1] Ershova A.S.,[2,3,4,*] Karyagina A.S.,[2,3,4] Spirin S.A.,[1,2,5] Alexeevski A.V. [1,2,5]
[1]*Lomonosov Moscow State University, Faculty of Bioengineering and Bioinformatics, Moscow 119992, Russia;*
[2]*Lomonosov Moscow State University, Belozersky Institute of Physical and Chemical Biology, Moscow 119992, Russia;*
[3]*Gamaleya Institute of Epidemiology and Microbiology, Moscow 123098, Russia;*
[4]*Institute of Agricultural Biotechnology, the Russian Academy of Sciences, Moscow 127550, Russia;*
[5]*Scientific Research Institute for System Studies, the Russian Academy of Science (NIISI RAS), Moscow 117281, Russia*
[*]*Corresponding author. asershova@gmail.com*

Key words: restriction-modification, restriction sites, bacteria, phage, site avoidance

Motivation and Aim: Restriction-modification (R-M) systems defend prokaryotes from invasion of foreign DNA recognizing and cleaving specific DNA sequences. For dozens of cases it was shown [1], that presence of an R-M system leads to the avoidance of its recognition site in the genomes of prokaryotes and their viruses. In this work a large scale analysis of thousands of currently available complete genomes was performed to estimate the press of R-M systems, encoded in prokaryotic genome, on the number of R-M systems recognition sites in genomes of prokaryotes and their viruses.

Methods and Algorithms: To study R-M system influence on the representation of the sites in genomes of prokaryotes and their viruses, method of estimation of underrepresentation suggested by Karlin et al. was used [2].

Results: We found that sites of R-M systems are underrepresented in genomes of prokaryotes or their viruses in a half of the cases. At the same time, the number of underrepresented potential R-M systems sites is three times greater than that can be explained by the encoded R-M systems. We also shown the sites of R-M systems recently acquired by a genome are often not avoided, while at least a part of underrepresented potential sites are the traces of lost R-M systems.

Conclusion: Representation of sites might shed light on the coevolution of R-M systems and genomes of prokaryotes and their viruses.

References:

1. E.P. Rocha, A. Danchin, A. Viari (2001) Evolutionary role of restriction/modification systems as revealed by comparative genome analysis // Genome Res, 11: 946-958.
2. S. Karlin, L.R. Cardon (1994) Computational DNA sequence analysis // Annu Rev Microbiol, 48: 619-654.

BIOLOGICAL GRAPH DATA BASE AND ITS APPLICATIONS

Ryasik A.A.*, Temlyakova E.A., Sorokin A.A.
Institute of Cell Biophysics RAS, Pushchino, Russia
e-mail: arc7an@gmail.com
**Corresponding author*

Key words: graph database, Neo4j, python, BLAST, NCBI

Motivation and Aim: Graph is the natural way to represent and analyze networks of biological objects and interactions among them. At the same time biological data are usually stored in relational databases and tabular files. We are going to bridge that gap by construction database of molecular biology data using graph-based storage system Neo4j [1]. The data for our database is going to be collected from the major authoritative well-known databases like UniProt for protein sequences, GenBank for genomic features, etc.

Results: We developed a Python module to communicate with Neo4j for biological data storage, update and retrieval. This module uses functions from BioPython to process molecular biology data and access to external databases. Communication with Neo4j backend is implemented with help of py2neo module. Infrastructure has functions for searching, writing, reading and mining data from a number of external databases (GenBank and UniProt). NCBI BLAST service is used to create and store sequence similarity links between elements of genomes.

Conclusion: The graph-based storage system for biological data has been developed together with python-based module, which contains a number of useful functions to fill, maintain and analyze molecular biology data within Neo4j graph database storage.

Availability: Source code is available at https://github.com/arc7an/BiomeDB.

Acknowledgements: This work was supported by RFBR grant 14-04-31793 mol_a.

References:

1. Neo4j web-site (Neo Technology): http://www.Neo4j.org/

DYNAMICS OF NONLINEAR CONFORMATIONAL EXCITATIONS IN FUNCTIONAL REGIONS OF pTTQ18 PLASMID

Ryasik A.A.*, Grinevich A.A., Yakushevich L.V.
Institute of Cell Biophysics RAS, Pushchino, Russia
e-mail: arc7an@gmail.com
**Corresponding author*

Key words: DNA, solitons, numerical methods, sine-Gordon equation, pTTQ18 plasmid sequence

Motivation and Aim: Study of the nonlinear DNA dynamics is an important and urgent problem of theoretical biophysics. It is well known that conformational mobility of DNA determines its basic functions, such as replication and transcription. However, the relation between the dynamical characteristics of the polynucleotide sequences and their biological activity is not clear enough. In this work we tried to find this relation in the studies of moving of conformational excitations (kinks) in the pTTQ18 plasmid sequence [1].

Methods and Algorithms: We used modified sine-Gordon equation with variable coefficients to describe the kink's movement in inhomogeneous DNA sequence, consisting of several inhomogeneous regions separated by boundaries. Then with the help of the method of concentrations [2] the equation coefficients were averaged over the functional regions of the sequence. It made possible to reduce the problem of the DNA kink dynamics to the problem of movement of the kink in an effective potential field defined as a set of wells and barriers, corresponding to the functional regions. The equation has been solved numerically using the finite difference schemes.

Results: The kink trajectories in the real sequence — the plasmid pTTQ18— have been obtained. We found several types of kink's behavior: reflection from the boundary, passing through the boundary with velocity increase or decrease. These types of behavior were found to be depended on the kink's initial velocity, initial position and dissipation coefficient values.

Conclusion: The results show that there is a relation between the arrangement of the functional regions in the sequence and DNA dynamical properties.

Availability: The Matlab code is available on request.

References:

1. PTTQ18.GB, http://www.lifesci.dundee.ac.uk/groups/mike_stark/pTTQ18_DNA.txt
2. Yakushevich L.V., Krasnobaeva L.A. (2008) International Journal of Non-Linear Mechanics 43(10):1074-1081.

GRAPH MODEL OF TYPE I DIABETES

Ryzhkov P.A.*, Ryzhkova N.S.
Scientific Research Institute of Biology SFEDU, Rostov-on-Don, Russia
e-mail: rygkov-p@yandex.ru
**Corresponding author*

Key words: type I diabetes, IDDM, graph theory, web-application, gene network, HLA

Motivation and aim: Type I diabetes (also called IDDM, Insulin Dependent Diabetes Mellitus) is one of the most common autoimmune disease in children. It is a complex polygenic disorder characterized by full β-cell loss due to autoreactive T-cell attack. Despite multiple studies the detailed mechanism of type I diabetes development still remains unclear. In this study we collect data of different types relating to T1D development and try to incorporate it into one vivid and simple mathematical model.

Methods and Algorithms: We used Action Script 3.0 to visualize data from MySQL database on web-based application. The data obtained from literature was introduced into database manually via PHP-based administrative system. Some network properties of the model are calculated as described by Pavlopoulos [1] (cluster analysis) and by Mason and Verwoerd [2] (centrality measures).

Results: Web-based graph model of type I diabetes have been developed. The model consist of i) database having specific structure for storing data about nodes and edges of the model and corresponding information needed for convenient visualization and analysis of the data presented; ii) database administrative system intended for introducing and correction of information about the nodes and their relationships, said information is base for graph model of T1D; and iii) software module for data visualization and analysis having possibility to scale the model and show information about selected model components.

Conclusion: The developed graph model allows analyzing different factors involved in the development of T1D and relationships thereof in the frame of single cross functional model by convenient visualization and presence of analytic functions including calculation of different metrics and cluster analysis.

Availability: Available on request from the authors.

References:

1. Pavlopoulos G.A. et al. (2011) Using graph theory to analyze biological networks, *BioData Min.*, **28**: 4:10.
2. Mason O., Verwoerd M. (2007) Graph theory and networks in Biology, *IET Syst Biol.*, **1(2)**:89-119.

COMPUTER ANALYSIS OF HUMAN SNP CONTAINING SITES BY METHODS OF TEXT COMPLEXITY ESTIMATIONS

Safronova N.S.[1]*, Orlov Y.L.[1, 2]
[1]*Novosibirsk State University, Novosibirsk, Russia;*
[2]*Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia*
e-mail: taschasafronova@bionet.nsc.ru
**Corresponding author*

Key words: SNP, mutations, human genome, text complexity, nucleotide poly-track

Motivation and Aim: Analysis of genomic sequences issues the challenge to search for the regions with low text complexity, which could be functionally important in genomes. It could affect structural characteristics of DNA sequences and decrease accessibility of such genome regions for high-throughput sequencing. The methods of nucleotide sequence complexity estimation [1] have wide area of application; also they can be effectively used for DNA analysis on the whole genome, short genome regions and for analysis of SNP in special. SNPs (single nucleotide polymorphisms) are very important in studying human diseases. This work continues the work started earlier [2]. It's devoted to the using the developed program complex with new module in SNP research in human chromosome. For the analysis there were taken long regions [-100;+100] nt on the flanks around the mutation position and short regions [-10;+10] nt on the flanks from UCSC Genome Browser. Also we used available SNP data from rat and mouse genomes (http://www.ncbi.nlm.nih.gov/SNP/)

Methods and Algorithms: Firstly the short regions were studied, and for them there was admitted a certain type of changes of the complexity near the mutation, so secondly the long regions were studied to find possible random changes of complexity. We studied the complexity of these genome sequences by different measures with sliding window. We used Lempel-Ziv complexity, linguistic complexity, Shannon's entropy and monomers frequency measures. Then we studied the nucleotide composition, analyzed the frequency of oligonucleotides occurrence by lengths 5-7 and built the profiles of occurrence for the most frequent oligonucleotides.

Results: Studying the complexity showed that in the mutation region there was a special character of complexity observed: before the mutation the complexity profile got visibly lower, then followed the hop and then the complexity fell again. So the graph of the complexity in this region looked like pit-peak-pit. The composition analysis showed that regions of mutation are saturated by A and T nucleotides. Further frequency analysis showed that there's a significant saturation with poly-A and poly-T tracks.

Conclusion: Neighboring-nucleotide effects on mutation rate can be revealed by text complexity estimations. Human and mouse genome SNP containing sites have similar complexity profiles.

References:

1. Y.L. Orlov, V.N.Potapov (2004) Complexity: an internet resource for analysis of DNA sequence complexity. Nucleic Acids Res. 32(Web Server issue):W628-33.
2. P. Putta et al. (2011) Relatively conserved common short sequences in transcription factor binding sites and miRNA, *Vavilov journal of genetics and breeding,* **15** (4): 750-756.

GRAPH DATABASE FOR MOLECULAR BIOLOGY: ADVANTAGES OF THE GRAPH REPRESENTATION OF DATA

Sergeev A.V.[1]*, Temlyakova E.A.[2], Sorokin A.A.[2]
[1]*Institute of Mathematical Problems of Biology RAS, Pushchino, Russia;*
[2]*Institute of Cell Biophysics RAS, Pushchino, Russia*
e-mail: antonsergeevphd@gmail.com
**Corresponding author*

Key words: graph database, data integration, database structure, and molecular biology data representation

Motivation and Aim: There are a great number of knowledge-specific databases that store essential and detailed information about a particular organism, types of reactions, molecular mechanisms and objects, etc. Most of that data stored in tabular form either in local files or in relational databases. On the contrary, many aspects of modern biology and medicine are naturally mapped to a graph or network structures. Another important aspect is a data structure mismatch when two databases represent the same biological object differently. This makes a fast and reliable data integration system an urgently needed.

Methods and Algorithms: We suggest using graph representation of molecular biology data taken from major authoritative well-known databases (UniProt, GenBank, RegulonDB, etc). All the data are stored as a one colored attributed graph with molecular objects as nodes with predetermined types of relations and object-specific properties. Data from different sources could be stored in different nodes, so let end user (not the database designer) to decide which data source have to be used. The highly flexible and easily scaled graph-orientated storage system Neo4j [1] serves as a backend and query system for the database.

Results: The prototype of graph database contained information about *Escherichia coli* genome, proteome, metabolome, transcriptional and translational regulation, etc is constructed. The key design issues and decisions are described and their consequences are analyzed.

Conclusion: Graph representation of molecular biology data provides new opportunities for queering database, information retrieval and analyzing patterns in data.

Availability: The first version will be available soon.

Acknowledgements: This work was supported by RFBR grant 14-04-31793 mol_a.

References:

1. Neo4j web-site (Neo Technology): http://www.neo4j.org/

THE PROGRESS IN PHYSICAL MAPPING OF CHROMOSOME 5B OF BREAD WHEAT *TRITICUM AESTIVUM*

Sergeeva E.M.[1], Timonova E.M.[1], Bildanova L.L.[1], Koltunova M.K.[1], Nesterov M.A.[1], Magni F.[2], Frenkel Z.[3], Dolezel J.[4], Faris J.[5], Sourdille P.[6], Feuillet C.[7], Salina E.A.[1]*
[1]*Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia;*
[2]*Institute of Applied Genomics, Udine, Italy;*
[3]*University of Haifa, Haifa, Israel;*
[4]*Institute of Experimental Botany, Olomouc, Czech Republic;*
[5]*USDA-ARS Cereal Crops Research Unit, Fargo, USA;*
[6]*UMR INRA-UBP Genetics, Diversity and Ecophysiology of Cereals, Clermont-Ferrand, France;*
[7]*Bayer Crop Science, Morrisville, USA*
e-mail: salina@bionet.nsc.ru
**Corresponding author*

Key words: *Triticum aestivum*, physical mapping, BAC-library, SSR-markers, ISBP-markers

Motivation and Aim: The bread wheat *Triticum aestivum* is an important crop and has a complexe genomic structure. *T.aestivum* is an allohexaploid (2n = 6x = 42, genomic formula BBAADD), and has the genome size of $17*10^9$ bp, more that 80% of which consists of repetitive sequences. The high standard sequencing of individual wheat chromosomes is in progress. The key to successful chromosome sequencing is the obtaining of physical map of chromosome. Our aim is to develop the physical map for wheat chromosome 5B which bears a number of agronomically important genes such as *Vrn*, S*kr*, *Ne*1, *Ph*1; and has a length nearly 870 Mbp: the short arm (5BS) is of 290 Mbp and the long arm (5BL) is of 580 Mbp.

Methods and Algorithms: The chromosome-specific BAC-libraries (Bacterial Artificial Chromosome) obtained for both arms. BAC clones were fingerprinted with SNaPshot method, and assembled using to contigs using LTC program. For contig anchoring to the chromosome we constructed the genetic and cytogenetic maps, the BAC-anchoring done by PCR-screening with mapped markers.

Results: The 5BS library consists of 43776 BAC-clones and 5BL library of 76 800 clones, that corresponds to approximately 15-fold coverage. For 5BS we performed the fingerprinting (restriction mapping) of BAC-clones that was consequently assembled into 277 contigs contained 26094 clones. The essential stage for contig anchoring to chromosome 5B is the construction of genetic and cytogenetic maps. For genetic mapping we used the F2 population consisting of 366 individual plants obtained from crossing of Chinese Spring (CS) cultivar with the disomic substitution line CS-5B *dic* where 5B chromosome of CS replaced by the 5B chromosome from *Triticum dicoccoides*. For the saturation of maps we tested for parents' polymorphism the 113 publicly available SSR markers and 660 newly developed on the base of 454-sequences ISBP markers. At the present moment the 30 polymorphic SSRs and 50 ISBPs are included to genetic map of 5B. Also 51 SSRs and 194 ISBPs were assigned to specific chromosomal BIN on 5B cytogenetic map; the markers to 5 known 5B genes (*Skr*, *TaCBFIIIc-B10*, *Vrn-B1*, *Chi-B1* and *Ph1*) included to map. 45 of mapped markers were anchored to contigs from 5BS BAC-library.

Conclusion: Here, we demonstrate the approach to construction of physical maps of individual chromosomes of complex genomes such as wheat which will be useful to accelerate map-based cloning, gain new insights into genome evolution, and provide a foundation for reference sequencing.

THE STRUCTURAL ORGANIZATION AND EVOLUTION OF 5S rDNA OF WHEAT CHROMOSOME 5BS BY DATA OF PARTIAL SEQUENCING

Sergeeva E.M.*, Koltunova M.K., Afonnikov D.A., Vasiliev G.V., Salina E.A.
Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
e-mail: sergeeva@bionet.nsc.ru
**Corresponding author*

Key words: 5S rDNA, allopolyploid, wheat, 5BS chromosome, IonTorrent, BAC-clone

Motivation and Aim: The genes coding for 5S ribosomal RNA (5S rDNA) are presented as clusters of tandemly repeated units and are the popular object for cytogenetic, molecular and phylogenetic studies in plant genomes. In allopolyploid wheat and its relatives it was demonstrated the wide range of genomic reorganizations involving 5S rDNA which mechanisms are still unclear [1-3]. To further explore the mechanisms affecting the 5S rDNA clusters during genome evolution it is necessary to obtain the 5S rDNA-spanning genomic sequences. The sequencing of individual chromosomes of common wheat *Triticum aestivum* ($2n = 6x = 42$, genomic formula BBAADD) is in progress. The aim of our study is the isolation and sequencing of genomic sequences spanning the 5S rDNA clusters located on the short arm of wheat chromosome 5B (5BS). The subsequent analysis includes: (1) the comparison of 5S rDNA-spanning sequences to those from the genomes of other cereal species; (2) the development of new molecular genetic markers to wheat 5S rDNA loci and their tracking in wheat and its relative species.

Methods and Algorithms: The isolation of genomic sequences containing 5S rDNA performed by PCR-screening of 5BS chromosome-specific BAC-library. BAC sequencing done using IonTorrent method. Sequence assembling *de novo* performed by MIRA program. Annotation of DNA sequences performed using BLAST algorithm with NCBI and TREP databases. The PCR markers designed with Primer3 program.

Results: The two BAC clones 010O13 and 025F09 containing 5S rDNA were isolated from 43776 BAC clones of 5BS-chromosome specific BAC library. We identified the four spanning regions with length from 2 to 11 kbp that represent (1) uncharacterized genomic sequences; (2) LTR-retrotransposon *Laura*. The comparison with publicly available 5S rDNA-spanning genomic sequences was performed. The 3 new molecular markers to 5S rDNA were developed and tested on a set of wheat species.

Conclusion: 5S rDNA loci are frequently associated with retrotransposons which confirms previous suggestion about the involvement of retroelements in 5S rDNA genomic rearrangements [4]. The use of data on the regions spanning the 5S rDNA is useful in the studies of evolution of these loci.

References:

1. O. Raskina et al. (2004) Quantum speciation in *Aegilops*: Molecular cytogenetic evidence from rDNA cluster variability in natural populations. *PNAS* **101(41)**:14818-23.
2. A.B. Shcherban' et al. (2008) Analysis of 5S rDNA changes in synthetic allopolyploids *Triticum* × *Aegilops*. *Mol Biol (Mosk*) **42(4)**:604-11.
3. B.R. Baum, M. Feldman. (2010) Elimination of 5S DNA unit classes in newly formed allopolyploids of the genera *Aegilops* and *Triticum*. *Genome* **53(6)**:430-8.
4. R. Kalendar et al. (2008) Cassandra retrotransposons carry independently transcribed 5S RNA. *PNAS* **105(15):**5833-8.

PUTATIVE CANDIDATE GENES TRANSCRIPTIONALLY UPREGULATED BOTH IN ACUTE AND CHRONIC PHASE OF RESPONSE DURING INFESTATION OF MICE AND SYRIAN GOLDEN HAMSTERS WITH LIVER FLUKES OF OPISTHORCHIIDAE FAMILY

Shamanina M.Y.
Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
e-mail: marinash@bionet.nsc.ru

Key words: host-parasite interaction; *Opisthorchis viverrini*; *Clonorchis sinensis*; mouse; hamster; transcription

Motivation and Aim: Understanding of mechanisms underlying host-parasite interactions is a crucial item in preventing parasitic invasions and early diagnostics of precancerous conditions that could be caused by parasites. Infestations with *Opisthorchis viverrini* (*Ov*) or *Clonorchis sinensis* (*Cs*) (Trematoda: Opisthorchiidae) are known to induce biliary epithelium fibrosis and to promote cholangiocarcinoma development [1]. Generally acute (AP) and chronic (CP) phases are distinguished in clinical manifestation of host-parasite response. The aim of this study was to detect differentially expressed genes in both AP and CP.

Methods and Algorithms: Array data annotated in GSE15447 (FVB/N male mice; *Cs*; 1 week after infection) [1] from Gene Expression Omnibus [2] were analyzed using GEO2R tool [3]. Generated "Top 250" gene selection was compared with the set of genes obtained by Wu et al (Syrian hamsters; *Ov*; 1, 3, 6 months after infection) [4].

Results: It was determined that during host-parasite response five genes, namely *Ltf*, *Anxa2*, *Mfge8*, *Socs3*, *Krt8*, were transcriptionally upregulated both in AP and CP.

Conclusion: As a rule, gene expression in CP is paid more attention than that in AP [1, 4]. Experimental verification of the prediction made, in combination with functional annotation, will allow to reconstruct the time course of transcription in AP and to clarify the role of AP in modulating CP.

Acknowledgements: This work was supported by Budget project "Molecular mechanisms of pathological processes: role of Opisthorchiidae in carcinogenesis" VI.60.1.1. State Registration Number 01201280335.

References:

1. D.M. Kim et al. (2009) Gene expression profiling in mouse liver infected with Clonorchis sinensis metacercariae, *Parasitol Res*, **106**:269–278.
2. Edgar R., Domrachev M., Lash A.E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository, *Nucleic Acids Res*, **30**:207-10.
3. T. Barrett et al. (2013) NCBI GEO: archive for functional genomics data sets – update, *Nucleic Acids Res*, **41**(Database issue):D991-995.
4. Z. Wu et al. (2011) Candidate genes involving in tumorigenesis of cholangiocarcinoma induced by Opisthorchis viverrini infection, *Parasitol Res*, **109**:657–673.

METHODS OF DETECTION AND ESTIMATION OF EVOLUTIONARY CONSERVED ELEMENTS OF RNA SECONDARY STRUCTURE

Sherbakov D.Yu.[1]*, Adelshin R.V.[2], Anikin A.S.[3], Gornov A.Yu.[3]
[1]*Limnological Institute SB RAS, Irkutsk, Russia;*
[2]*Irkutsk Antiplague Research Institute of Siberia and Far East, Russia;*
[3]*Institute for System Dynamics and Control Theory, Irkutsk, Russia*
e-mail: sherb@lin.irk.ru
**Corresponding author*

Key words: doublet model, conserved hairpins

Motivation and aim: Functions of RNA are not limited to the transfer of linearly coded information. There are several programs calculating the most probable folding of RNA molecules, but although they improve their performance continuously, it appears that for sufficiently long sequences of RNA global energetic optimum for one single best fold does not exist. On the other hand, if long enough random sequence of nucleotides is generated and then the optimum folding into hairpins is calculated, large proportion of nucleotides will be paired. Thus, large faction of hairpins predicted may meaningless.

In case a hairpin possesses a physiological role, base substitutions must preserve secondary structure. Molecular evolution if constrained by RNA secondary structure is described by a doublet model, thus the search for functional hairpins turns into the search for sequence stretches having significant bias in doublet substitution frequencies. Here we report the set of programs for detection of conserved hairpins in DNA or RNA sequence alignments using a folding model and bayesian testing of the hypothesis that a doublet model fits better then the best single nucleotide model.

Algorythms and Methods: The nucleotides involved into formation of a hairpin almost never form a single continuous string, thus it is important to split a long sequence into smaller fragments so that they could be examined one by one. We divided the sequence so that all pairs (doublets) within each fragment are complete. Only fragments with the numbers of pairs larger then certain threshold are taken into account. Unpaired positions at this stage are dropped. This method depends critically on the model of secondary structure accepted.

Nucleotide sequences are recoded into doublets and few values describing doublet diversity are calculated. All predictions of RNA secondary structure were made using Vienna RNA package. Statistical significance of the advantage of the doublet model over the best single nucleotide one was estimated using Bayes factor.] This was automated by program sstems taking as the input the alignment and the model of it's folding in dot-bracket notation.

In order to obtain various sorts of a reference datasets in suitable formats using mixed doublet and single nucleotide models, the simulation program null_H is written. It evolves a random nucleotide sequence along a random tree using mixed doublet and single nu- cleotide models. The doublet model may be made imperfect by allowing certain amount of mutations towards stem-destroying doublets.

Results: Two approaches towards finding evolutionary conserved elements of RNA secondary structure are developed and tested with simulated and natural data sets. It is possible to use this as the tool for "fast and dirty" search for the parts of an alignment evolving according to a doublet model and thus likely to contain an element of some functional significance. The speed and efficiency of this method allows one to sets several alternative folding models to quite large data set. Statistical comparison between a doublet model and single nucleotide model is possible.

Availability: All programs and scripts are available from http://sherb.lin.irk.ru/programs.html

Acknowlegments: A.G. and A.A. were partially supported by RFBR grant 13-01-00470

UVA-INDUCED MODIFICATIONS OF LENS ALPHA-CRYSTALLIN

Sherin P.S.[1]*, Zelentsova E.A.[1, 2], Sormacheva E.D.[1, 2], Duzhak T.G.[1], Tsentalovich Yu.P.[1, 2]
[1]*International Tomography Center SB RAS, Novosibirsk, Russia;*
[2]*Novosibirsk State University, Novosibirsk, Russia*
e-mail: sherin@tomo.nsc.ru
**Corresponding author*

Key words: UVA irradiation, kynurenic acid, alpha-crystallin, radical reactions, aggregation, oxidation

Motivation and Aim: UVA (320-400 nm) is the most intense part of sun UV irradiation, reaching the surface of our planet. This irradiation is believed to be involved into the lens normal aging and the cataract development. The main chromophors of the human lens are kynurenine and its derivatives, acting as UV-filters. Under UV light these molecules exhibit low yield of reactive species (triplets), which can react with surrounding proteins. With age, the accumulation of photoinduced modifications might lead to the irreversible changes of protein structure such as coloration, loss of solubility and aggregation. The aim of this work is to study the nature of protein modifications induced by photoexcited UV-filters.

Methods: Alpha crystallin (bovine) solution, containing 6M urea, was irradiated (355 nm) under anaerobic conditions in the presence of kynurenic acid (KA). The mechanisms of reactions of KA triplet state with proteins were studied by Laser Flash Photolysis; the photolysed proteins were analyzed by gel-electrophoresis and HPLC-MS/MS. The same experiments were performed in the presence of lens intrinsic antioxidants, ascorbic acid (Asc) or glutathione (GSH), which concentrations were close to natural conditions (few mM).

Results: KA in the triplet state reacts with alpha crystallins via tryptophan and tyrosine residues, yielding the corresponding radicals. Subsequent radical reactions result in the formation of large (> 200 kDa) water-insoluble and urea-SDS-resistant aggregates, which are formed via cross-linking of crystallin monomers. Other observed modifications include loss of tryptophan/tyrosine residues, oxidation of methionine-1 and formation of unknown product of tryptophan-9 absorbing at 340 nm. In the presence of Asc none of such modifications was observed; however, in the presence of GSH only the aggregation was blocked. This is explained by effective quenching of triplet KA by Asc and radical scavenging by GSH.

Conclusion: Under UVA irradiation the intrinsic UV-filters might induce irreversible modifications of alpha crystallins including protein aggregation, oxidation of methionine and chemical modifications of tryptophan and tyrosine residues. The ascorbic acid is effective antioxidants preventing the major part of photoinduced modifications of lens proteins.

Acknowledgements: This research was supported by RFBR (Projects 14-03-00027, 14-03-00453 and 14-03-31189), by the Ministry of Education and Science RF, and by the Division of Chemistry of RAS.

INTEGRATED GENOME-ORIENTED INFORMATION SYSTEM FOR MONITORING AND CONTROL OF BIOLOGICAL SYSTEMS

Shlikht A.G.*, Kramorenko N.V.
Far-Eastern Federal University, Vladivostok, Russia
e-mail: schliht@mail.ru

Key words: genomics, proteomics, metabolomics, dynamic systems, expert systems

Motivation and Aim: The traditional page-based Internet data-genome organization does not provide adequate performance and the complexity of the analysis of the information stored on the global portals. The aim of the work is the creation of a specialized tool information system (IS) support integrated research in genomics, transcriptomics, exomics, proteomics, metabolomics (GTPM), integrated with traditional models of dynamic systems.

Methods and Algorithms: In the framework developed by the IS is based on the system-cybernetic structured model [1], based on the classical theory of dynamic systems and database and knowledgebase technologies

Results: Developed an integrated genome-based IS monitoring and management of living systems. Storage and access to information (genome, transcriptome, exome, proteome, metabolome) is not in the traditional horizontal text and vertically in the form of indexed entries in the database table up to the nucleotide, amino acid, metabolite. Such an approach, losing in memory offers excellent performance and data mining by means of data and knowledge bases. IS more integrated with the system of monitoring the state of the environment and human health [2], which enables it to meet the challenges of the GENOTYPE-HABITAT-PHENOTYPE. IS through a hyperlink associated with the world informational resources (NCBI, EBI, KEGG, etc.)

Conclusion: The instrumental IS can serve as tools for the creation of a specialized Bank GTPM, and modeling of dynamic processes in biological systems.

References:

1. Shlikht A.G., Kramorenko N.V. (2014) Gene-centric system-cybernetic structured model of a living organism. The international conference Biotechnology and quality of Life, 338-340.March, 18-20, 2014, Moscow, Russia.
2. А.Г. Шлихт, Н.В. Краморенко. (2007) Автоматизация процессов монитринга и управления состоянием окружающей среды и здоровья человека.-Владивосток: Изд-во ТГЭУ, 2007.-324с.

ELoE – A WEB-APPLICATION FOR ESTIMATION OF GENE TRANSLATION ELONGATION EFFICIENCY IN VARIOUS ORGANISMS

Sokolov V.S.*, Zuraev B.S., Lashin S.A., Matushkin Yu.G.
Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia;
Novosibirsk State University, Novosibirsk, Russia
e-mail: sokovlad1@bionet.nsc.ru
**Corresponding author*

Key words: translation elongation efficiency, codon composition, mRNA secondary structures, web application

Motivation and Aim: Expression efficiency is one of major characteristics describing genes. To measure or estimate it is a great challenge of modern biology. Unfortunately, experimental data estimating the efficiency of gene transcription or translation are available only for small groups of organisms. Bioinformatical approaches allow to solve this problem (at a first approximation) via the analysis of genes nucleotide sequences. In this study, we present a special web-application EloE (Elongation Efficiency). The application calculates an efficiency of translation elongation for all genes and then sorts them in descending order.

Methods and Algorithms: EloE program performs the analysis of genes nucleotide sequences. The key parameters of the analysis are gene codon composition and presence of perfect local inverted repeats in mRNA [1].

Results: The developed software allows carrying out the analysis of genomes for several organisms. Individual directory with all results of calculations is created for each organism. User has an opportunity to change default parameters of calculation and mark the results he want to save for more flexible control of the program runs. Organisms' genes sorted in descending order by calculated translation elongation efficiency, plots with distribution of perfect local inverted repeats, plots with comparison of translation elongation efficiency of genes with the same names in different organisms, summary table with all organisms and parameters of its genomes and other results are available after completion of program calculation.

Conclusion: The software can be useful in preliminary estimation of translation elongation efficiency of genes in organisms, which experimental data are not available yet. Some results can be used, for instance, in modeling of artificial genetic structures in genetically engineered experiments.

Availability: http://www.bionet.nsc.ru/razrabotki/prikladnyie-razrabotki/programmyi-dlya-evm.html.

Acknowledgements: This work was supported in part by the program of the Presidium of Russian Academy of Sciences "Origin and evolution of the biosphere" (no. 15).

References:

1. V.A. Likhoshvai, Y.G. Matushkin. Differentiation of single-cell organisms according to elongation stages crucial for gene expression efficacy // FEBS. 2002.

VARIATION OF ELONGATION EFFICIENCY INDEX OF ARCHAEA GENES DURING EVOLUTION

Sokolov V.S.*, Gunbin K.V., Matushkin Yu.G.
Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
e-mail: sokovlad1@bionet.nsc.ru
**Corresponding author*

Key words: mRNA secondary structures, translation efficiency, evolution, Archaea

Motivation and Aim: Various organisms are known to optimize primary structure of genes in different ways to increase its translation efficiency [1]. Some organisms optimize codon composition, others optimize count and/or stability of mRNA secondary structures. In the majority of Archaea species the number of secondary structures in mRNA plays a crucial role in determining translation elongation efficiency. The main goal of this study is an analysis of a primary structure of ancestral and present nucleotide sequences of Archaea genes to identify its significant changes/optimizations, which can lead to increase of elongation translation rate.

Methods and Algorithms: Ancestral nucleotide sequences of Archaea genes were reconstructed taking into account the reconstructed protein sequences. Elongation efficiency indexes (EEI2) were calculated for all nucleotide sequences (ancestral and present-day) by a special program. The index value depends on count of secondary structures in mRNA [1, 2]. The results were placed in nodes of reconstructed phylogenetic trees.

Results: As a result of analysis of phylogenetic trees genes with significant changes in EEI2 values were found. These changes are due to significant increase or decrease of amount of mRNA secondary structures. Generalization of results from analysis of individual genes shows that the most considerable variations of EEI2 are associated with sharp changes of bacteria's ecology. Thus, the most significant rearrangements in primary structure of genes happen when organism changes its habitat. It indicates an adaptation to a new environment.

Conclusion: The results of this study confirm that Archaea optimize primary structure of their genes to increase the efficiency of elongation translation. It was shown that the organism environment is one of the most important factors in forcing the optimization of genes primary structure.

Availability: The special software will be available on the server of ICG soon (http://www.bionet.nsc.ru/razrabotki/prikladnyie-razrabotki/programmyi-dlya-evm.html).

References:

1. V.A. Likhoshvai, Y.G. Matushkin. Differentiation of single-cell organisms according to elongation stages crucial for gene expression efficacy // FEBS Lett. 2002. V. 516 (1-3). P. 87-92.
2. V.S. Sokolov, V.A. Likhoshvai, Yu.G. Matushkin. Gene expression and secondary mRNA structures in different *Mycoplasma* species // Vavilov Society of Geneticists and Breeders. 2013. V. 17 (4/1), P. 639-650.

COMPUTER ANALYSIS OF HUMAN GENE EXPRESSION DATA USING BioGPS DATABASE OF MICROARRAY AFFYMETRIX U133

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

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

Motivation and Aim: Analysis of gene expression using microarrays is associated with many problems in bioinformatics. The most common method is the technology of synthesis of short oligonucleotide probes on a surface of the microchip developed by the Affymetrix, which allow to simultaneously measure the expression of thousands of genes. . Microchips GeneChip U133 series are widely used in clinical practice, large amount of experimental data was collected (BioGPS (biogps.org/), GEO NCBI databases). Initial design of oligonucleotide probes microchip may not correspond to a target transcript (gene) and contain a number of technical problems associated with genomic annotation of samples. It leads to inconsistent results, so for working with data expression it is required to develop a specific software package that allow to filter noisy expression signals on the microchip and make it easier to work with large amount of data, and aim of work is to create a set of programs for multipurpose computer research on expression of genes on the microchips.

Methods and Algorithms: To meet issues of human genome analysis of expression data was developed set of computer programs to work with existing databases. Tools in C++ for processing Affymetrix U133 microarray data, including algorithms for estimating coefficients of correlation and filtration for samples of genes, are implemented to perform the different tasks, e.g. to identify characteristics of genes which actively expressed in human brain tissues and analyze characteristics of the expression of pairs of transcripts which co-localized in the genome including cis-antisense transcripts.

Results: Program for analyzing correlation was developed, which is also useful for analyzing human brain gene expression and gene networks (e.g. were obtained correlation matrix for networks of cholesterol level regulation and circadian rhytm regulation), for research microchip quality [2], and analyzing tissue specificity (group of gene with high expression in brain cells was identified using BioGPS DB [1]), including 3-dimensional chromosomal contacts.

Conclusion: Among samples microchip Affymetrix U133, that are represented in the BioGPS database, were identified some with high expression, collected samples of genes for which the expression is higher in brain structures, created a tool that allows to create a chart for correlations of expression of gene pairs. Structural features of genes with high expression (number of exons, the length of the transcript, link with alternative splicing) were detected. It is planning to integrate this tool into a package for statistical data processing expression of genes that is developed in ICG SB RAS.

Availability: https://github.com/amspitsina/gene-expr, http://about.me/anastasiaspitsina

References:

1. I.V. Medvedeva, O.V. Vishnevsky, N.S. Safronova, O.S. Kozhevnikova, M.A. Genaev, A.V. Kochetov, D.A. Afonnikov, Y.L. Orlov (2013) Computer analysis of data on gene expression in brain cells obtained by microarray tests and high-throughput sequencing // Vavilov Journal of Genetics and Breeding, V.17, № 4/1, p. 629-638.
2. Orlov Yu.L. et al (2007) Quality assessment of the Affymetrix U133A&B probesets by target sequence mapping and expression data analysis. In Silico Biol. 7(3):241-60.

WHOLE GENOME ANALYSIS OF A-to-I RNA EDITING USING SINGLE MOLECULE SEQUENCING IN *DROSOPHILA*

StLaurent G.[1, 2]*, Shtokalo D.[2, 3, 5], Tackett M.R.[2], Nechkin S.[2, 3, 5], Antonets D.[4], Vyatkin Y.[2, 5]*, Savva Y.A.[1], Kapranov P.[2], Lawrence C.E.[1], Reenan R.A.[1]
[1]*Brown University, Providence, RI, 02912, USA;*
[2]*St. Laurent Institute, One Kendall Square, Cambridge, MA 02139, USA;*
[3]*A.P. Ershov Institute of Informatics Systems SB RAS, Novosibirsk, Russia;*
[4]*State Research Center of Virology and Biotechnology «Vector», Novosibirsk, Koltsovo, Russia;*
[5]*AcademGene LLC, Novosibirsk, Russia*
e-mail: georgest98@yahoo.com, vyatkin@gmail.com
**Corresponding authors*

Key words: ADAR, post-transcript modification, NGS, Machine Learning, *Drosophila*

Motivation and Aim: RNA editing is a mechanism of post-transcriptional modification of RNA sequence. Numerous examples exist where RNA editing changes the function of the RNA and also protein encoded by it. Editing mediated by ADAR's (adenosine deaminases that act on RNA) is one of the most common editing mechanisms in eukaryotes. As such, detection of authentic A-to-I RNA editing events produced by these enzymes has generated significant interest in the genomics community. It has also presented a major challenge considering the numerous confounding events which obscure the actual signal. Here, we present a robust RNA editing pipeline based on a combination of single-molecule sequencing (SMS), an exhaustive validation step, and the application of machine learning methods that allow us to almost completely annotate the inosinome of an adult organism – male *Drosophila*.

Methods and Algorithms: Using RNA-Seq data obtained for fruit flies over 13 million locations in the *Drosophila* genome were found to have at least one aligned read with an A-to-G substitution on either strand, which correspond to almost 1/3 of all adenosine positions in the genome. To distinguish between real editing sites and the many types of False Positive events, all the sites containing reads with A-to-G substitutions were subjected to sophisticated filtering procedure which resulted in ~43K candidate sites with validation rates well below 20%. We tested a number of machine learning classification algorithms to distinguish between True Positives and True Negatives and the Random Forest (RF) approach demonstrated the best performance. Application of RF model to the entire 43K candidate sites resulted in its separation into three domains Tier 1 (2.5K sites), Tier 2 (14K), and Tier 3 (26.5K) according to predicted RF score. A Sanger validation of a set of randomly selected candidate sites from Tier 1 and Tier 2 resulted in a validation rate of 75% and 21% respectively.

Results and Conclusion: Specifically we have identified what we estimate to be ≥80% of the editing sites in male *Drosophila*, which allows us to draw general conclusions about the biological properties and functions of RNA editing in *Drosophila*. We show that most editing occurs in non-coding portions of the transcriptome. Sites within mRNAs show a 5' to 3' bias and are significantly enriched for non-synonymous recoding events in highly conserved regions. Sites of ADAR modification exhibit minor sequence motif conservation and tend to cluster within transcripts. Finally, we provide evidence that editing globally regulates expression levels of edited mRNAs and that presence of RNA editing is strongly positively correlated with alternative splicing.

LOGICAL MODELLING OF Nanog-DEPENDED TRANSCRIPTIONAL GENE NETWORK OF EMBYONIC CARCINOMA STEM CELLS

Stepanenko I.L.*, Ivanisenko V.A.
Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
e-mail: stepan@bionet.nsc.ru
**Corresponding author*

Key words: embryonic stem cells, boolean model, gene networks, pluripotency, self-renewal

Motivation and Aim: Embryonic stem cells are self-renewing, pluripotent cell lines, characterized by their potential to differentiate into all cell types. Transcription factor Nanog is critical for maintaining the self-renewal and proliferative properties of embryonic stem cells (ESCs). We have reconstructed the Nanog regulated gene network and have analyzed their topology and dynamical properties. Boolean modeling was used to describe the behavior of the gene network. Studies of local network topology and dynamics can be used to investigate as well to predict the novel markers for ES cells.

Methods and Algorithms: We use an *in silico* approach to identify novel potential regulators of ES regulatory networks by using gene perturbation experiments together with qualitative modeling. To reconstruct and analyze associative gene networks of Nanog-depended genes was used ANDSystem software package. We have developed Boolean model for Nanog regulated gene network that assume only two states (ON or OFF) for each component. The Boolean rule for each node is determined based on the nature of interactions between the nodes (activation or inhibition). This rule can be expressed using the logical operators AND, OR and NOT.

Results: P19 embryonal carcinoma stem cells, which have properties of both ES and EC cells, have been used to identify potential targets of Nanog (Choi SC et al., 2012). It has been shown that knockdown or overexpression of Nanog in P19 cells affect the expression of the genes involved in p53- and cell cycle-signaling pathways. We developed a Boolean network model for P19 embryonal carcinoma stem cells. The literature based Nanog regulated gene network includes 21 nodes and 56 edges. We used SQUAD methodology to convert a signaling network into a dynamical system, even in the total absence of kinetic information (Cara AD et al., 2007). Using SQUAD we have simulated the network dynamics and the perturbation of Nanog-regulated Boolean network. Dynamical analysis allowed us to enumerate the two steady states and mutation simulation to identify the role of the specific nodes within the network (Jdp2, CycE1, Tob1, Bcl6 and *etc.*) that regulate oxidative stress, cell cycle and p53- signaling pathway network modules.

Conclusion: This model can be useful to test the coherence of experimental data and to hypothesize gene interactions that remain to be discovered.

Acknowledgements: The work was supported by budget grant VI.61.1.2.

IS THE SINGLE CELL CHIP-Seq TECHNIQUE POSSIBLE?

Subkhankulova T.*, Naumenko F.**
*Imperial College London_*
e-mail: tsubkhan@imperial.ac.uk
**Corresponding author*

Key words: single cell, ChIP-Seq, modelling, amplification

Motivation and Aims: Chromatin immunoprecipitation followed by high-throughput DNA sequencing allows high resolution, genome-wide location analysis. Currently, conventional methods of ChIP-Seq require chromatin from thousands, sometimes millions cells per analysis, although there is growing demand in techniques to test as little as a few cells or even a single cell. This is particularly essential in developmental studies, early prenatal and clinical diagnostics, cancer research. Indeed, scaling the ChIP-Seq techniques down to a single cell levels permits to obtain valuable information concerning epigenetic mechanisms in individual cells, and, therefore, gives insight in fine tuning in gene regulation which results in normal development or, on the contrary, pathological processes. However, technical complexity, absence micro-devices and appropriate controls to test epigenetic events at a single cell levels, and high costs of experiments makes single cell location analysis a challenging and even unachievable goal. This encouraged us to perform the simulation of ChIP-Seq analysis *in silico* and confirm the most significant conclusions using experimental data.

Materials and methods: To run simulations and analysis of simulated ChIP-Seq data High performance computing suit (HPC), Imperial College London, was used. Experimental results were obtained using MethylCollector Ultra Kit (Active motive Ltd) followed by quantitative PCR to assess enrichment of loci of interests.

Results: We created algorithm closely resembling the ChIP-Seq experimental procedure, which allowed varying the experimental parameters such as number of cells taken, the percentage of signal loss and background contamination, chromatin fragmentation, fragment size selection and additional amplification.

We have tested the influence of "number of cells", the levels of "signal loss" or "background contamination", and combination of these factors on resultant WIG and BED-files, which represented the outcome of *in silico* ChIP-Seq experiments. Here, we determined that single cell-ChIP may produce poor data even when the signal loss and background contamination were minimal or literarily absent; which is, of course, very unlikely in real experimental circumstances. The striking observation was that if we applied amplification procedure prior the "size selection" and "library construction" we practically restored majority of "peaks", or the regions where our "putative factor" interacted with "chromatin". Even for single cell data with "signal loss" up to 10 percent and "background contamination" up to 10 per sent we still provide the relatively good level of coincidence with original "pattern" file. The optimization of the fragmentation size and "omitting of size selection" also improved the similarity between initial pattern and resultant output files.

To prove our findings we applied optimized ChIP parameters to perform similar to ChIP technique (allowing selection of methylated CpG island fragments) starting only with 100 or 10 cells. We discovered that DNA fragments associated with methylated CpG islands were enriched in positive loci comparably to negative unmethylated fragments in 40 and 10 folds, correspondently.

Conclusions: The results of our study confirmed that single cell location analysis is possible, although the technique requires further improvements and optimizations. Researchers, perhaps, need to adopt the nano- and micro-technological modifications such as microfluidics and look for stronger (e.g., covalent) interactions to select chromatin: protein complexes. Alongside with dropping of sequencing costs it may make the single cell epigenetic analysis a routine procedure.

**The ChIP-Seq simulation software (ishIP) is available on request.

EVOLUTION OF MODERN HUMAN AND RECOMBINATION OF MEMES

Suslov V.V., Ponomarenko M.P., Gunbin K.V.
Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
e-mail: valya@bionet.nsc.ru

Key words: hominids, modern human, ape, brain, evolution, memes, TATA-box

Motivation: Traditionally the emergency of modern humans (MH) explained by the widening of exchange and recombination of memes (ERM - "when ideas have sex" [1]) and selection of genes facilitating to this exchange[1]. These features would have to accelerate the brain evolution and lead to an increase of society openness, accompanied by the culture progress. [1]. Per contra: genome projects have shown that the brain is the most conservative MH's organ, and brain expressing positive selected genes typically are expressed not only brain [2]. Progress correlate with society openness only ~300 years (after the formation of global trade networks), earlier, the progress flashed in civilization-semiisolates[2].

Results: It was found that TATA-boxes of neuron expressing genes of ancient hominids had practically equal activity to TATA-boxes of MH. On the contrary, modern apes have TATA-boxes with significant changes of its activity. Hence, the origin of MH does not related with major neuroanatomic change but with small number of sequential variations aimed at change in brain operations without changing its energy metabolism. This is consistent with the genome-wide data and antropometric data. According the genome-wide data mostly positive selected genes are ones of the main service infrastructures (MSI - energy and ions metabolism, stress, etc.) working in any tissues. According the antropometric data a gradual increase in brain size does not correlate with the culture.

Conclusion: MH brain evolved under the principle of advanced evolution of MSI in inter- and intragenic levels. Only after these any neuroanatomical changes make adaptive sense. Similarly only after formed cultural *MSI* (F.Braudel's structures of everyday life – SEL [5]) in civilization-semiisolates ERM make adaptive sense.

Acknowledgements: Budget Project VI.61.1.2., IP Presidium RAS No.28 «BOE» .

References:

1. Ridley M. The Red Queen. NY, Harper Perennial. 2003. 416 p.
2. Khaitovich P. et al. (2005) // Science. 309, 5742. P. 1850–1854.
3. The problem of race in the Russian Anthropology / Ed. T.I. Alekseeva, L.T. Yablonsky. M, IEA RAS. 2002. 96 p. (in Russ.).
4. Lordkipanidze D. et al. (2013) // Science. 342, 6156. P. 326–331.
5. Braudel F. A History of Civilizations. NY, Penguin Books, 1995. 640 p.

[1] In other words, small and may be neutral genetic change must be driven by cultural change and vice versa (Baldwin effect) [2]. About the one's limitations see thesis "Baldwin effect as evolution silencer".

[2] For archanthrops and early neanthrops races are not typical. Only MH have races which conflicts [3,4] with the conception of their wide migration activity. The phenomenon of "centers of civilization" is traced from classical Cro-Magnon through "hydraulic civilization" of the ancient East to ancient Greece and medieval Europe. Any "center of civilization" was formed in spatially confined complex landscape or a combination of landscapes. It is interesting to note that Vavilov's centers of diversity have similar features, and in such conditions the *MSI* (radical) is formed most rapidly.

THE GENOMIC TEXT CHARACTERISTICS AND GC CONTENT ARE RELATED TO THE BACTERIAL GENOME EVOLUTION

Suslov V.V.[1]*, Safronova N.S.[1, 2], Orlov Y.L.[1, 2], Afonnikov D.A.[1, 2]
[1]*Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia;*
[2]*Novosibirsk State University, Novosibirsk, Russia*
e-mail: valya@bionet.nsc.ru
**Corresponding author*

Key words: evolution, prokaryotic genomics, GC content, non-specific adaptation

Motivation and Aim: GC content, genome size, and other genome characteristics are integral genome features limiting potential spectrum of licenses (habitats) of prokaryotic organisms. There are other specific and non-specific adaptations which are important for evolution. Most studied non-specific adaptations are inducible, such as stress response. Traditionally, GC content is treated as non-inducible non-specific adaptation maintaining genome stability. In average, the more habitat agents damaging DNA or codon-anticodon interactions are, the higher is GC content of the given genome. At extreme habitat environments this trend lost due to other specific molecular mechanisms maintaining genome stability and gene expression. But Wu et al. [1], studying eubacteria, assume that GC content is semineutral trait, related to features of replication/reparation machinery in different prokaryotics taxa (dimeric combinations of DNA polymerase III alpha subunits: dnaE1/ dnaE2/dnaE3 groups).

Methods and Algorithms: We have downloaded from NCB ftp-site 1214 annotated and 1586 assembled complete genome sequences of prokaryotic organisms [2]. We counted correlation of GC content and genome size for all genomes and for groups – 3 archeal taxa and joint archeal group, and 19 bacterial taxa and 1 joint eubacterial group.

Results: Correlations of genome size and GC content are high for Archaea (104 species, r=0,35), Eubacteria (1478 species, r=0,59) and all Prokaryota (r=0,46). The same correlation trend is present after separation of the genome sample by taxa for all Archaea, and most Eubacteria excluding Deinococcus-thermus, Fusobacteria and Planctomycetes. Trend for Eubacteria agreed with Pol III classification in [1], but for 3 exceptions. We argue that for Deinococcus-thermus and Planctomycetes specific GC content maintained adaptively in opposite to replication/reparation features. Comparing genome size, GC content and preferable habitats we show that for eurybiont species genome size is larger, but for complex ecosystems (such soil environment) GC content is larger because extremophiles, highly specialized prokaryotes don't hold high GC content. In particular only ecology grouping "prokaryotes of soil ecosystems" is characterized in average by both high GC content and large genome size.

References:

1. H. Wu et al. (2012) On the molecular mechanism of GC content variation among eubacterial genomes, *Biol Direct,* 7**:** 2.
2. V.V. Suslov, D.A. Afonnikov, N.L. Podkolodny, Yu.L. Orlov (2013) Genome features and GC content in prokaryotic genomes in connection with environmental evolution. *Paleontological Journal*, 47(9), 1056-1060.

SSRFace: AN IDENTIFICATION AND SEARCH TOOL FOR GENOMIC AND TRANSCRIPTOMIC SSR

Tan H., Hou X.*
State Key Laboratory of Crop Genetics and Germplasm Enhancement, Key Laboratory of Biology and Germplasm Enhancement of Horticultural Crops in East China, College of Horticulture, Nanjing Agricultural University, Nanjing, P.R. China
e-mail: tanhuawei1991@163.com
**Corresponding author*

Key words: SSR, identification, genomic, transcriptomic

Motivation and Aim: Simple sequence repeats (SSRs) have proven to be highly polymorphic, easily reproducible, co-dominant markers. The advent of the genomics age has resulted in the increasingly massive amounts of available genomic and transcriptomic DNA sequence data, while the existing development tools are dispersive and inconvenient.

Results: Here, we developed an identification tool for genomic and transcriptomic SSR. After DNA submission by users, our combined pipeline would hunt for SSRs and design relevant primers, especially in transcriptomic analysis. Moreover, a SSR search interface containing over 20 plants was exploited in advance to search SSRs rapidly. For example, we identified 25682 SSRs in ~94Mb Arabidopsis thaliana genomic sequence and 3302 SSRs in its 35386 gene CDS sequences. We also identified 145842 and 4568 SSRs in Brassica rapa ~284Mb genomic and 41019 gene CDS sequences, respectively. The proportion of SSR unit sizes in B. rapa CDS was extremely unevenly distributed: 177 (3.87%) were mononucleotide, 195 (4.27%) dinucleotide, 4155 (90.96%) tri-nucleotide, 3 tetra- , 1 penta- and 37 hexa-nucleotide SSRs. In contrast, 100591 (68.97%) were mononucleotide, 33887 (23.24%) dinucleotide, 10051 (6.89%) tri-nucleotide, 962 tetra- , 190 penta- and 161 hexa-nucleotide SSRs in B. rapa genomic sequence. A similar distribution also exists in other plants.

Conclusion: We designed an identification and search tool for genomic and transcriptomic SSR development. Plant genomic sequences contain much more SSRs than transcriptomic sequences. Interestingly, mononucleotide and tri-nucleotide were the most abundant type founded in plant CDS sequences and genomic sequences, respectively.

Availability: SSRFace tool and plant SSRs are available at: http://nhccdata.njau.edu.cn/SSRFace .

GRAPH ANALYSIS OF *E. COLI* TRANSCRIPTION REGULATION

Temlyakova E.A.*, Sorokin A.A.
Institute of Cell Biophysics RAS, Pushchino, Russia
e-mail: evgenia.teml@gmail.com
**Corresponding author*

Key words: transcription regulation, graph analysis, *E.coli*, RegulonDB

Motivation and Aim: We have constructed a graph database containing information about *Escherichia coli* genome, proteome, metabolome, transcription and translation regulation, etc. All the data were taken from external well-known databases (Genbank, Uniprot, RegulonDB, etc). To show advantages of data storage in graph structures, we analyzed *E.coli* transcription regulation using subgraph based upon well-known RegulonDB data.

Methods and Algorithms: Experimental and analytical data from RegulonDB [1] were used to create a graph structure that was stored in Neo4j graph database [2]. The analysis was performed using Cypher queering language and Python.

Results: Based on RegulonDB data about *E.coli* transcription regulation we construct a colored attribute graph with 13 nodes (operon, transcription unit, gene, promoter, TFBS, TF, etc) and 8 relations (contains, encodes, binds, initiates, etc) types. Each node contains node-specific properties for analysis and reference to original data source. A few specific regulation patterns were revealed and discussed.

Conclusion: Graph representation of molecular biology data allows researchers to overview data at various scales: starting from an object characteristics and up to whole network of a particular mechanism in an organism.

Availability: Available upon request.

Acknowledgements: This work was supported by RFBR grant 14-04-31793 mol_a.

References:

1. H. Salgado et al. (2012) RegulonDB (version 8.0): Omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more, *Nucleic Acids Research,* **41(D1):** D203-D213.
2. Neo4j web-site (Neo Technology): http://www.neo4j.org/

RECONSTRUCTION OF ASSOCIATIVE GENE NETWORKS SPECIFIC TO TARGET BIOLOGICAL PROCESSES AND PHENOTYPIC TRAITS

Tiys E.S., Demenkov P.S., Saik O.V., Popik O.V., Ivanisenko V.A.*
Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
e-mail: salix@bionet.nsc.ru
**Corresponding author*

Key words: Associative gene networks reconstruction, connectivity, text-mining

Motivation and Aim: Large amounts of information on the interactions between molecular genetic objects, complex structural-functional units of an organism and phenotypic traits are accumulated in the scientific literature and knowledge bases and require automatic data collection, formalization and analysis. One approach to solve this problem is the reconstruction of associative gene networks, which includes two tasks: (1) Evaluation of the specificity of the reconstructed network to the target biological process or phenotypic trait. (2) Finding the optimal expansion of a gene network from the set of genes that are experimentally associated with the target biological process and serve as initial genes (IG) for the reconstruction process.

Methods and Algorithms: In this study the ANDSystem tool [1] was used for the reconstruction of the associative gene networks. ANDSystem consists of network visualizer ANDVisio and knowledge base ANDCell, containing information extracted by text-mining from PubMed abstracts and databases.

To solve the first problem (1), we proposed a method for assessing the specificity of the reconstructed network, based on comparison of its connectivity with networks based on IG randomly selected from all the genes of the organism under study. Network connectivity depends on the availability of information about the interactions of genes. In order to take into account study bias due to well-studied IG, the restrictions on the degree of randomly selected vertices were imposed. The second way to assess the specificity of the reconstructed network is to compare the representation of connected pairs of genes within the group of IG with their surrounding network.

To solve the problem (2) of network expansion it is proposed to use thresholds for adding vertices by their specificity and centrality. Specificity is evaluated as the ratio of the number of IG associated with a given vertex, to the total number of vertices associated with a given vertex in the entire network of the organism. Centrality of the vertex is calculated as the number of IG associated with it, divided by the total number of IG.

Results: Using the developed approach we reconstructed associative gene networks for IG taken from the KEGG PATHWAY database. Comparison of connectivity of these networks with those constructed from randomly selected genes showed a significant difference between them. Thus, the viability of this approach was demonstrated.

Acknowledgements: This work was financially supported by project VI.61.1.2

References:

1. P.S. Demenkov et al. (2011) ANDVisio: a new tool for graphic visualization and analysis of literature mined associative gene networks in the ANDSystem, In silico biology, 11:149-161.

INCREASING THE NUMBER OF PARALOGS FOR ENZYMES INVOLVED IN TRYPTOPHAN BIOSYNTHESIS DURING THE EVOLUTION OF LAND PLANTS

Turnaev I.I.[1]*, Gunbin K.V.[1], Akberdin I.R.[1], Mironova V.V.[1, 2], Omelyanchuk N.A.[1], Afonnikov D.A.[1, 2]
Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
e-mail: turn@bionet.nsc.ru
**Corresponding author*

Motivation and Aim: An increase of complexity of organisms by increasing the number of cell types occurred during the evolution of land plants. The amino acid tryptophan is not only a necessary component of the protein sequences, but it is also a precursor of a number of compounds for defense against pathogens and pests, as well as auxin, a key hormone of plant morphogenesis. Paralogues are genes that evolved through duplication. Our purpose was to study the functional significance of duplications of genes encoding tryptophan biosynthetic enzymes, leading to an increase in the number of paralogues in the evolution of land plants.

Methods and Algorithms: The samples of ASA/ASB, TRP, PAI, IGPS, TSA, TSB protein sequences of tryptophan biosynthesis enzymes were formed using a search for homologs by BLASTP v. 3.10 tool in 24 plant sequenced genomes (5 - green algae, 1 - moss, 1 - lycopodium, 4 - monocots and 13 - dicots). Protein alignment was performed by Promals3D. For more accurate identification of homologues in the sequences of protein families, we reconstructed phylogenetic network using method ProteinMLdist/ NeighborNet from the package SplitsTree4 (v. 4.12.8). Correlation analysis was performed using Statistica package.

Results: In this study we evaluated the relationship between the complexity of organisms of different plant taxa and the number of paralogues for tryptophan biosynthetic enzymes in them. To assess the complexity of organisms we used the index F_{TAPs} (TAPs - transcription associated proteins). F_{TAPs} - the ratio of proteins associated with transcription to the total number of proteins in the genome [1]. F_{TAPs} can be accurately estimated from genome-wide data and correlates well with the number of cell types as the characteristic of the organism complexity [1]. Analysis of the relationship between plant F_{TAPs} and a total number of paralogues of enzymes in the tryptophan biosynthetic pathway revealed a statistically significant positive correlation ($p=0.0005$). Also F_{TAPs} showed significant correlations with the numbers of paralogues for ASA ($p=0.0025$), ASB ($p=0.031$), PAI ($p=0.0033$), IGPS ($p=0.015$).

Conclusion: In evolution, plants adapted to the changeable environmental conditions at conquest of land by increasing the complexity of their morphology, such as appearance of roots, leaves and stems in ferns, horsetails and club mosses, the transition from spores to seed on the seed ferns or the emergence of flowers and fruits in angiosperms. The detected correlation may be explained by the fact that both the increase in complexity of the morphology of organisms and the increase of the number of paralogs tryptophan synthesis enzymes are adaptations to increased variability of land environmental conditions.

Acknowledgements: This study was supported by Programs RAN Presidium 6.6 and № 28, grant 5278.2012.4 and 11-04-01748-a, Integration Projects SB RAS № 39 and 130.

References:

1. Lang D., Weiche B., Timmerhaus G. et al. (2010), Genome Biol. Evol., 2, 488–50.

COMPUTER AND EXPERIMENTAL ANALYSIS OF MOLECULAR MECHANISMS OF GENE EXPRESSION REGULATION IN BRAIN TUMOR CELLS

Vasiliev G.V.[1]*, Gubanova N.V.[1], Spitsina A.M.[2], Safronova N.S.[2], Orlov Y.L.[1, 2]
[1]Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia;
[2]Novosibirsk State University, Novosibirsk, Russia
e-mail: genn@bionet.nsc.ru
**Corresponding author*

Key words: gene expression, microarray, glioma, gene networks, human genome

Motivation and Aim: Analysis of gene expression in brain tumors is of great importance for research of cancer treatment. Hypothesis of specific tumor stem cells in glioma has wide acceptation and should be tested on independent data using complementary computer and experimental methods. We apply computer approaches to analyze gene expression data on microarrays and high-throughput transcriptome sequencing data. Aim is to reveal specific gene expression patterns (signature) characterizing tumor aggressiveness for diagnostics and decease prediction. Interactions between gene and gene products and reconstructed gene networks were studied using STRING database. We used available data from GEO NCBI and develop own experimental models of cell lines derived from glioma patients.

Methods and Algorithms: We downloaded available clinical gene expression data from GEO (GDS1962) taken from patients with varying tumor grades (I, II, III and IV). Computer program in C++ was developed set to work with existing databases, process Affymetrix GeneChip microarray data. We used quality control estimates for probe sets of Affymetrix microarrays described earlier [1]. Then we studied differential expression to reveal sets of differentially expressed genes. In addition we studied correlations between gene expression levels across patients' samples. Using the algorithm developed earlier we calculate correlation coefficients for all the genes selected and random control samples.

Results: Strong positive correlation coefficients between genes from the selected gene sets were shown. It indicates that pattern of gene expression specific for cancer has complex nature, and the genes work together in the gene network, forming regulatory contours and activating one another. We used set of genes differentially expressed in brain tissues found as tissue-specific in BioGPS database (biogps.org/) as control to reveal tumor-specific genes in glioma. We plan use other gene expression data sources such as Cancer Gene Atlas, Allen Brain Atlas.

Acknowledgements: The work is supported by RFBR (14-04-01906).

References:

1. Orlov Y.L. et al. (2007) Quality assessment of the Affymetrix U133A&B probesets by target sequence mapping and expression data analysis. *In Silico Biol*. 7(3):241-60.

ANALYSIS OF THE DEGENERATE MOTIFS IN REGIONS OF BINDING SITES OF TRANSCRIPTION FACTORS ESSENTIAL FOR EMBRYONIC STEM CELLS MAINTENANCE

Vishnevsky O.V.*
Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia;
Novosibirsk State University, Novosibirsk, Russia
e-mail: oleg@bionet.nsc.ru
**Corresponding author*

Key words: embryonic stem cell, degenerate oligonucleotide motifs, transcription factor binding sites

Motivation and Aim: Despite the critical roles of transcriptional regulators in the embryonic stem cells maintenance, detailed knowledge of their targets is lacking. A lot is unclear about the interactions of transcription factors affecting target genes in embryonic stem cells. Clarification of these issues is necessary for understanding the molecular nature of pluripotency, self-renewal, and reprogramming. New high-performance computer approaches are needed to handle the new massive experimental data.

Methods and Algorithms: The sets of non-overlapping [-100;100] regions relative to the experimentally proved binding sites of 13 sequence-specific mouse TFs (Nanog, Oct4, STAT3, Smad1, Sox2, Zfx, c-Myc, n-Myc, Klf4, Esrrb, Tcfcp2l1, E2f1, and CTCF) are formed of the sequences, derived in ChIP-Seq experiment, in region [-100;100] relative to positions of maximum peak height. A new full-exhaustive computer program has been developed for detection of degenerate region-specific oligonucleotide motifs. It is based on the high-performance GPU technologies to process the massive data.

To compare the effectiveness of the heuristic and full-exhaustive search methods for detection of the signal, were generated samples of random sequences of varied length containing a motif with a given level of degeneracy and located in different number of sequences. Info-gibbs program was used to identify the motives heuristically. The method compares well with existing methods like MEME, BioProspector, Gibbs or GAME on both synthetic and biological data sets. The similarity between a motif obtained and the sample motif was measured with the Kullback-Leiber distance.

Results: It has been shown that such heuristics as Info-gibbs work well for relatively short sequences. With increasing of the length of the sequences up to several hundred nucleotides quality of their performance drops noticeably, this required the use of full-exhaustive approach. For each of the samples were obtained sets of significant motives. The analysis of co-localization of motifs and their classification using TRRD and TRANSFAC DB allowed identifying the potential composite elements. It was shown that some of them correspond to known elements such as Sox-Oct. Another part may correspond to the binding sites of other transcription factors, or some physico-chemical characteristics of the nucleotide context.

Conclusion: Using a new high-performance computer system allowed us to obtain new information about the molecular mechanisms of functioning of transcription factors and to identify new contextual signals which may be subject to further experimental analysis.

Acknowledgements: The work was supported by the SB RAS project 136.

SEARCH FOR ALTERNATIVE TRANSLATION STARTS IN THE GENOME OF *MYCOBACTERIUM TUBERCULOSIS*

Vishnevsky O.V.*, Kochetov A.V.
Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia;
Novosibirsk State University, Novosibirsk, Russia
e-mail: oleg@bionet.nsc.ru
**Corresponding author*

Key words: translation start sites, open reading frames, tuberculosis, degenerate oligonucleotide motifs, Markov model

Motivation and Aim: The existing methods of the translation start sites prediction usually based on the evolutionary conservatism of start sites and open reading frames. This approach does not allow identifying of the orphan genes, arising in the evolution *de novo*, and not having homologues even in closely related species. In addition, a substantial number of prokaryotic organisms have not yet sequenced or annotated genomes of closely related species. This compels researchers to use as homologues genes of evolutionarily distant species, leading to erroneous annotation of start sites. Furthermore, such techniques impose restrictions on the minimum size of the open reading frame, increasing errors. In particular, it was shown that some short open reading frames corresponds to the experimentally determined peptides, ranging in size from 16 to 50 amino acids and has been demonstrated their expression *in vivo*. Some of these peptides are localized in the membrane.

Methods and Algorithms: We have developed a new method of analysis and recognition of potential open reading frames and start sites. It is based on the method of detection of degenerate oligonucleotide motifs of fixed length and Markov chain model. Computer program that uses high-performance GPU, based on the CUDA technology has been developed. A set of 1583 sequences in the region [-50, 50] with respect to the translation start sites of H37Rv genome *Mycobacterium tuberculosis*, non-overlapping with other genes was analyzed. For training of a Markov model were used the open reading frames of corresponding start sites.

Results: 41 reliable degenerate oligonucleotide motifs were identified during the analysis of areas of *M.tuberculosis* translation start sites. Only a small portion of them corresponded to known Shine-Dalgarno box and (A/G)UG codon and localized in the area of the translation start. This can be explained by the high complexity of organization of the translation start regions. Furthermore, except canonical AUG represented in 60.9% of start sites, *M.tuberculosis* uses as starting GUG (33.5%) and UUG (4.5%) codons. It is interesting that the greatest significance and representation had distal motifs, located at a distance 30-40 bp. from the start of translation.

Conclusion: With the use of motives found and calculated Markov model in the genome of *M.tuberculosis* have been identified alternative translation start sites and the open reading frames. Further it is planned functional annotation of identified motifs and open reading frames.

Acknowledgements: The work was supported by the SB RAS project 136.

CONTROL OF THE miRNA PATHWAYS BY THE SECONDARY STRUCTURE AND ITS ACCOUNT IN THE PREDICTION TOOLS

Vorozheykin P.S.[1]*, Titov I.I.[1, 2]
[1]*Novosibirsk State University, Novosibirsk, Russia;*
[2]*Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia*
e-mail: pavel.vorozheykin@gmail.com
**Corresponding author*

Key words: miRNA, miRNA prediction, miRNA biogenesis, secondary structure, review

Motivation and Aim: MicroRNAs (miRNAs) are an abundant class of small non-coding RNAs that primarily influence gene expression either by the mRNA cleavage or by the translation arrest. The identification of novel miRNAs has become an important approach towards understanding of post-transcriptional gene regulation, on the other hand a revealing of the new miRNA features provides a possibility to predict the novel miRNAs and to maximize the prediction quality. Recent research has shown that the account of the secondary structure is important for the miRNA computer prediction.

Results and Conclusion: In this presentation we start from considering the three miRNA maturation processes – canonical, *Dicer*- and *Drosha*-independent. For each biogenesis we investigate the role of the secondary structure in the miRNA pathways. The common property which we have found is the excessive loop frequency in the miRNA anchor region and near the miRNA boundaries. Using recent experimental findings we show how the bulges and internal loops in the anchor region control the precursor strand selection (miR or miR*) and the miRNA sorting in *Ago*-protein families. Besides, the secondary structure near the boundaries could significantly affect the miRNAs' shifts which are important for duplex transfers, for the miRNA enzymes processing and for the forming miRNA's «seed» region.

Finally we review the existing computational miRNA prediction tools. We compare the tool performance, utility, speed and contribution of the secondary structure account to the prediction quality.

Acknowledgement: The work was supported by the RAS Presidium Program #6 "Molecular and Cell Biology".

ENHANCING ANTICANCER ACTIVITY OF DOXORUBICIN BY MOLECULAR IODINE COMPLEXES WITH BIOACTIVE ORGANIC LIGANDS AND LITHIUM SALTS

Yuldasheva G.A.[1], Ilin A.I.[1], Zhidomirov G.M.[2]
[1]*Scientific Centre of Anti Infection Drugs, Kazakhstan;*
[2]*Boreskov Institute of Catalysis SB RAS, Russia*
e-mail: yuldasheva57@rambler.ru; ilin_ai@mail.ru; zhidomirov@mail.ru

Key words: doxorubicin, iodine complex, DFT-method, MP2-method

Investigation of the combined effect of a medicinal product containing LiCl (I)- I_2-α-dextrin peptide complexes, and doxorubicin on the growth of Ehrlich ascites carcinoma (EAC) revealed that both intraperitoneal and oral administration of the above product in low doses resulted in a significant increase in the antitumor activity. The increase was more pronounced when using the medicinal product containing LiCl (I)- I_2-α-dextrin-peptide complexes, and doxorubicin.

Previously, we have demonstrated that the interaction of complexes LiCl (I)- I_2-α-dextrin peptide complexes with DNA nucleotides and can displace peptides and form a complex with molecular iodine coordinated by complexes of lithium halide with dextrin (complexes LiCl (I)- I_2-α- dextrin - nucleotide).

Inhibitors of topoisomerase I bind in one complex both onco-DNA nucleotide pairs and amino acid residues that are part of the active site of human DNA topoisomerase IB which is an enzyme that modifies and controls the state of cellular DNA.

DFT/B3PW91/6-31G ** shows that under the influence of LiCl (I)- I_2-α-dextrin-peptide complexes a structure is formed in onco-DNA in which the LiCl (I)- I_2-α-dextrin complex becomes the center of inhibition for both the onco-DNA nucleotides and the amino acid residue of arginine which is part of the topo I active center.

Doxorubicin is an anticancer preparation of the anthracycline group of antibiotics. Chemotherapeutic effect of doxorubicin is due to its ability to intercalate into DNA structure. According to the results of experimental studies the energy of stacking interactions of the antibiotic with a pair of nucleotide bases correlates with its anticancer activity.

The ability of doxorubicin to intercalate into DNA and form a structure in which the molecule of doxorubicin may be located parallel to the nucleotide pair and bound to it by the stacking interaction was investigated in the framework of RHF / midi and MP2/midi methods.

Our calculations indicate that the action of one and the same nucleotide pair is exposed to doxorubicin and LiCl (I) α-I_2-dextrin complexes the latter increases the energy of stacking interactions of doxorubicin with the nucleotide pair.

Thus, the effect of enhancing the anticancer activity of doxorubicin by molecular iodine complexes with and lithium halides and bio-organic ligands is caused by the formation of a structure in which the nucleotide pair is bound by stacking interaction with doxorubicin and coordination bond with LiCl (I) α-I_2--dextrin. In this structure, LiCl (I)-α-I_2-dextrin increases stacking interactions of doxorubicin with the nucleotide pair and becomes a center for inhibiting the amino acid residue of arginine which is part of the topo I active center.

LITHIUM HALIDES ENHANCE THE ANTICANCER ACTIVITY OF CISPLATIN

Yuldasheva G.A.[1], Ilin A.I.[1], Zhidomirov G.M.[2]
[1]*Scientific Centre of Anti Infection Drugs, Kazakhstan;*
[2]*Boreskov Institute of Catalysis SB RAS, Russia*

Key words: Ehrlich ascite carcinoma, lithium halides, DFT-method

An experimental investigation of the antitumor activity of preparations containing iodine complex with bio-organic ligands, and potassium and lithium halides (MeCl (I)-I_2-dextrin-peptide). It is shown that the combined action of the preparation containing complexes LiCl (I)- I_2-dextrin-peptide and cisplatin on the growth of Ehrlich ascite carcinoma (EAC) where the preparation is administered after the administration of cisplatin significantly enhanced the antitumor activity of the latter.

In [1] the ab initio quantum chemical method was used to calculate the structures that simulate the interaction of cisplatin with guanine bases of DNA. It was shown that the formation of stable complexes with guanine lead to the transition of guanine to the enol form or to the breakage of hydrogen bonds within the guanosine-cytosine nucleotide pair.

UV- and IR-spectra of the aqueous solution containing the LiCl(I)-I_2-dextrin-peptide complex and the interpretation thereof using the quantum chemical method DFT/B3PW91/midi indicate that inside the dextrin helix the following three complexes may be formed: a lithium halides complex with dextrin and peptides, a triiodide complex with dextrin, located inside the dextrin helix, and a complex of molecular iodine with halides of lithium and peptides.

Lithium halides contained in the medicinal product under study may form stable complexes with guanosine and the phosphate group. Complexation with guanosine at the five-membered ring nitrogen atom is more energetically profitable.

In the event that the product containing the LiCl (I)- I_2-dextrin peptide complex is administered prior to the administration of cisplatin the lithium halides form a stable complex with guanosine, and for that reason the number of guanosine bases able to freely interact with cisplatin decreases. This weakens the anticancer effects of cisplatin because the formation of a lithium halide-guanosine complex strengthens the hydrogen bond in the guanosine- cytosine pair. In the case of cisplatin when it is administered first, it forms a stable complex with guanosine, and only the phosphate group can be the site for coordinating lithium halides.

When interacting with the DNA topoisomerase I the active center of the enzyme interacts with the phosphate group. In vitro studies described in literature demonstrated that lithium chloride suppressed the growth of carcinoid cells. Our calculations indicate that one of the possible mechanisms for this effect could be the binding by lithium halides in one and the same complex of the phosphate group and the amino acid residues of arginine and tyrosine that are part of the active site of topoisomerase I.

Thus, the combined action of cisplatin and preparations containing LiCl (I)- I_2-dextrin peptide complexes forms a complex with guanosine, which brings about mutations in the DNA, while lithium halides inhibit the active centre of topoisomerase I. At the cellular level, these processes cause the disruption of transcription and replication, which leads to cell cycle arrest and apoptosis.

References:

1. I.L. Zilberberg, V.A. Avdeev, G.M. Zhidomirov //J.Mol.Struct.(Theochem) 1997. V.418. P.73-81.

CONTACT-BASED APPROACH TO STRUCTURAL CLASSIFICATION OF PROTEIN-DNA COMPLEXES

Zanegina O.N.[1]*, Karyagina A.S.[1, 3], Alexeevski A.V.[2, 3], Spirin S.A.[2, 3]
[1]*Belozersky Institute for Physical and Chemical Biology, Moscow State University, Moscow, Russia;*
[2]*Scientific Research Institute for System Studies (NIISI RAS), Moscow, Russia;*
[3]*Gamaleya Institute of Epidemiology and Microbiology, Moscow, Russia*
e-mail: zanolya@yandex.ru
**Corresponding author*

Key words: protein-DNA interactions, classification

Motivation and Aim: Systematization of protein-DNA complexes is necessary for understanding mechanisms of protein-DNA interaction. As the number of new structures rises every year, it is important to create an approach for automatic classification of new protein-DNA complexes. To consider both protein and DNA parts in interactions contact-based approach can be used in classification.

Methods: Tools of NPIDB [1], a database of protein-nucleic acid structures were used to to extract SCOP protein domains in contact with DNA. For each domain-DNA structure hydrogen bonds and hydrophobic contacts were listed. Each contact was assigned to secondary structure element of protein (helix, sheet, strand, turn) and DNA major/minor groove, sugar-phosphate backbone. For SCOP families interaction type was defined as a list of those contacts which were represented in all proteins.

Results: Protein-DNA contacts in 794 structures from 118 families of SCOP protein domains in contact with DNA were characterized. All variety of observed contacts was divided onto 37 interacting groups. In most cases several interacting groups were found within the same SCOP family. To reject crystallization artifacts in single structures and to find conservative protein-DNA contacts the procedure of automatic SCOP family classification were created. For 29 families contained 3 and more different proteins 11 interaction types were described.

Conclusion: Comparison of structures within domain families allow to distinguish conservative features. The variety of all observed contacts in family could help to predict additional contacts in a particular complex, which can appear due to a minor protein motions *in vitro*.

Acknowledgements: this work was partially supported by RFBR grant 13-07-00969.

References:

1. Kirsanov D.D., Zanegina O.N., Aksianov E.A., Spirin S.A., Karyagina A.S., Alexeevski A.V. (2013) NPIDB: Nucleic acid-Protein Interaction DataBase. *Nucleic Acids Res*. **41**(Database issue):D 517-23.

GLOBAL MAPPING OF PROTEIN UBIQUITYLATION WITHIN TNF-ALPHA SIGNALING PATHWAY USING GENEXPLAIN PLATFORM

Zhabereva A.S.[1, 2]*, Gainullin M.R.[1, 2], Ivanchenko M.V.[1], Kel A.E.[3]
[1]*Lobachevsky State University of Nizhny Novgorod, Nizhny Novgorod, Russia;*
[2]*Nizhny Novgorod State Medical Academy, Nizhny Novgorod, Russia;*
[3]*GeneXplain GmbH, Wolfenbüttel, Germany*
e-mail: zhabereva.anastasia@gmail.com
**Corresponding author*

Key words: transcriptional regulation, TNF-alpha signaling, NFkB-dependent gene expression, pathway analysis, geneXplain

Motivation and Aim: There are many evidences indicating that ubiquitylation controls activity of transcriptional factor as well as expression of particular genes. However, in system-wide analysis of signaling cascades ubiquitin-dependent regulation remains neglected. The present work aims for global mapping of ubiquitylated proteins, involved in TNF-alpha signaling pathway. Also, we try to estimate how ubiquitin-dependent degradation influence NFkappaB-dependent gene expression

Methods and Algorithms: Experimental transcriptomic data were downloaded from ArrayExpress (http://www.ebi.ac.uk/arrayexpress/). List of ubiquitylated proteins was compiled from recently published proteomic data sets and specialized ubiquitin-related databases. Reconstruction of TNF-alpha signaling pathway and modeling of NFkappaB-dependent gene expression was produced using WEB-version of geneXplain platform (http://platform.genexplain.com/bioumlweb/#), v.2.4.1.

Results: Based on analysis of differentially expressed genes, we reconstructed TNF-alpha signaling pathway and identified TNFalpha as the master regulator and NFkappaB as the common effector. At the next step, this pathway was enriched by information, reflecting protein ubiquitylation. Surprisingly, the majority of proteins, involved in TNF-alpha signaling, were identified as potential targets of modification by ubiquitin moiety. Further, we introduced parameter, corresponding to the rate of ubiquitin-dependent degradation of IkB, into dynamic model describing oscillations of nuclear form of NFkappaB. It was shown that both increase and decrease of ubiquitylation rate leads to dramatic effects on NFkB-dependent gene expression, in particular IkBα expression.

Conclusion: Sub-total ubiquitylation of proteins related to the TNF-alpha signaling pathway was demonstrated. Moreover, significant modulation of NFkB-dependent gene expression upon varied ubiquitin-dependent degradation rate was predicted. Thus, protein ubiquitylation belongs to widespread and essential regulatory events and should be considered by computational reconstruction of signaling/metabolic pathways.

Acknowledgements: This work is partially supported by the Russian Foundation for Basic Research (grants 12-04-01836, 14-04-01199) and Russian Federal Program (governmental contract 14.B37.21.1234).

5-METHYLCYTOSINE AND DNA OXIDATION: AT THE CROSSROADS OF EPIGENETICS, DNA DAMAGE, AND DNA REPAIR

Zharkov D.O.
Institute of Chemical Biology and Fundamental Medicine SB RAS, Novosibirsk, Russia
e-mail: dzharkov@niboch.nsc.ru

Key words: base excision repair, DNA glycosylases, epigenetics, 5-methylcytosine, 5-hydroxymethylcytosine, 5-hydroxymethyluracil, 8-oxoguanine

Motivation and Aim: 5-methylcytosine (mCyt) is a major epigenetic marker in DNA of higher eukaryotes. Recently, active demethylation of epigenetically regulated genes was shown to be dependent on base excision DNA repair (BER), which is normally responsible for removal of damaged bases from DNA to prevent mutagenesis, carcinogenesis, and cell death. BER is initiated by one of a series of enzymes called DNA glycosylases, which recognize and excise specific types of damaged bases. The abasic (AP) site that remains in DNA is then nicked by AP endonucleases, and firther processed by DNA polymerases and DNA ligases. Distinct DNA glycosylases specific for removal of mCyt or products of its oxidation and/or deamination have been found in cells from vertebrates and from plants.

Methods and Algorithms: We have analyzed several possible alternative pathways of active demethylation, in particular, a possibility of removal of mCyt during BER of the adjacent guanine-derived lesions.

Results: The major human glycosylase (OGG1) responsible for removal of 8-oxoguanine (oxoGua), an abundant oxidized purine, was able to initiate DNA repair in damaged CG-dinucleotides in which human AP endonuclease (APEX1) then removed mCyt by the virtue of its exonuclease activity. The cancer-associated OGG1 variant S326C was severely deficient in its ability to initiate this pathway and interact with APEX1. Interactions of mCyt and oxoGua thermodynamically destabilized CG-dinucleotides compared with their nonmethylated counterparts. Repair of oxidative damage in methylated $(CGG)_n$ runs, characteristic of fragile X syndrome, prompted DNA synthesis that could lead to the run expansion. Several human DNA glycosylases were characterized with respect to their activity towards mCyt and products of its oxidation and/or deamination, 5-hydroxymethylcytosine, 5-formylcytosine and 5-hydroxymethyluracil. We have also shown that plants possess a specific AP endonuclease-like protein, Ape1L, which is apparently optimized for the exonucleolytic removal of 3'-terminal residues.

Conclusion: Several possible pathways of active DNA demethylation may be operative in eukaryotic cells, thus providing a way for fine tuning of the cell's epigenetic state.

Acknowledgements: This work was funded by RAS Presidium (MKB 6.12).

STREPTOPHYTE ALGAE AND THE ORIGIN OF LAND PLANTS REVISITED USING THE CHLOROPLAST GENOMES AND NUCLEAR GENES

Zhong B.
Institute of Fundamental Sciences, Massey University, New Zealand
e-mail: b.zhong@massey.ac.nz

The phylogenetic branching order of the green algal groups that gave rise to land plants remains uncertain despite its fundamental importance to understanding plant evolution. To better understand the evolutionary history of land plants, we firstly analyzed a chloroplast genome data set including three new chloroplast genomes from streptophyte algae, and further applied a site pattern sorting method together with time-heterogeneous models to investigate the branching order among streptophytes and land plants. Next, Using recently available nuclear sequences from streptophyte algae, we applied the multispecies coalescent model to produce a congruent phylogeny that is robust to different nuclear data sets, in contrast to the conflicting phylogenies produced by the concatenation method. Our chloroplast and nuclear phylogenomic analyses support Zygnematales alone, or a clade consisting of Coleochaetales plus Zygnematales, as the closest living relatives of land plants. Furthermore, we suggest that the coalescent model can accommodate gene tree heterogeneity in deep-level phylogenies and can be potentially used to resolve other deep species phylogenies.

A MODEL OF TRICHOME SPACING PATTERN FORMATION ON GROWING WHEAT LEAF

Zubairova U.S.*, Nikolaev S.V., Doroshkov A.V., Afonnikov D.A.
Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
e-mail: ulyanochka@bionet.nsc.ru
**Corresponding author*

Key words: wheat leaf, trichome patterning, symplastic growth, computational model

Motivation and Aim: The epidermis of the wheat leaf is a single cell layer organized as longitudinally oriented parallel rows of cells that are elongated in the direction from base to leaf tip. Some of the epidermal cells differentiate into trichomes, the unicellular unbranched epidermal outgrowths. Trichomes form a specific spacing and size pattern on leaf blade surface; their size ranges from several mkm to millimeters. Different wheat cultivars characterized by different pubescence density and the trichome length distribution. Notably, trichomes in wheat arranged in rows at a similar distance from each other. We put forward a simple phenomenological model to account for this pattern formation on growing linear leaf blade.

Methods and Algorithms: We assume a simple mechanical cell-based model for symplastic growth of linear leaf containing the following steps: (1) in the direction of the leaf width generation of an initial cell layer from one cell due to cell growth and division, (2) in the direction base to leaf tip generation of parallel cell rows from every cell of the initial layer without either movement of the cells or new contacts between them and accompanied by mutual adjustment between all the cells. Also we assume that on the second step cells of initial layer produce some substances which diffuse along the cell rows and mark zones of stemness, amplification and elongation characterizing by its growth function and division rate.

Results: In the work hypotheses are tested that the trichome spacing pattern is established in amplification cell zone by the following mechanism selecting single trichome cell from otherwise equivalent epidermal cells. We propose that trichome produce a substance which diffuses along the cell row and inhibits the formation of other trichomes. We assume that the postulated inhibitor is destroyed by epidermal cells, so that a gradient is set up around a trichome cell, and that there is a threshold level of inhibitor below which development of the trichome cell begins (the above-threshold region defining an inhibitory zone). It is supposed that an epidermal cell embarks upon an irrevocable course of differentiation into trichome as soon as the concentration of inhibitor falls below threshold.

Conclusion: This model was used to fit experimental data on quantitative characteristics describing wheat leaf pubescence obtained by the method LHDetect2 (http://wheatdb.org/lhdetect2). We have optimized parameters of the model by fitting different patterns of trichomes distribution on the leaf blade for different wheat cultivars.

Availability: Program code and model parameters are available on request from the authors.

Acknowledgements: This work was supported by programs of the Presidium of the Russian Academy of Sciences nos. 6.6, and 28 (subprogram 2), SBRAS integration projects 47, 130.

A SIMPLE MECHANICAL CELL-BASED MODEL FOR SYMPLASTIC GROWTH OF LINEAR LEAF BLADE

Zubairova U.S.*, Nikolaev S.V.
Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
e-mail: ulyanochka@bionet.nsc.ru
**Corresponding author*

Key words: morphogenesis, leaf development, symplastic growth, epidermal patterning, computational model

Motivation and Aim: In grasses the developing leaf is established by parallel files of cells originating at the leaf base. The feature is symplastic growth where neighboring cell walls adhere and do not slide across each other. We developed a simple mechanical cell-based model for symplastic growth of linear leaf blade.

Methods and Algorithms: We assume an anisotropic symplastic growth model for linear leaf blade starting from a meristem-like layer of generative cells and then generating parallel cell rows from every cell of the initial layer. Each cell is characterized by its growing function, but growth oh the whole leaf blade is accompanied by mutual adjustment between all the cells: there are no movements of the cells or new contacts between them except cell division. Cells divide once they have reached a threshold area. Also we assume that some cells produce morphogens which diffuse along the cell rows and mark zones of stemness, amplification and elongation according to French-flag model [1]. For simulation of cell division and cell differentiation we used the formalism of parameterized L-systems [2].

Results and conclusion: The model developed as a Mathematica notebook can be used for building and running multi-cell virtual-tissue simulations to study cell patterns resulting from anisotropic symplastic growth of linear leaf blade.

Availability: Program code and model parameters are available on request from the authors.

Acknowledgements: This work was supported by programs of the Presidium of the Russian Academy of Sciences nos. 6.6, and 28 (subprogram 2), SBRAS integration projects 47, 130.

References:

1. Wolpert L. (1969) Positional information and the spatial pattern of cellular differentiation, J. Theor. Biol., 25 (1): 1–47.
2. Lindenmayer A. (1968) Mathematical models for cellular interaction in development, Parts I and II, J. Theor. Biol., 18: 280–315.

A DATABASE OF RHODOPHYTE PLASTID PROTEIN FAMILIES AND REGULATION OF moeB GENES

Zverkov O.A.*, Seliverstov A.V., Lyubetsky V.A.
Institute for Information Transmission Problems RAS (Kharkevich Institute), Moscow, Russia
e-mail: zverkov@iitp.ru
**Corresponding author*

Key words: orthology relationship, database of plastid protein families, plastids, algae, transcription regulation, molybdopterin biosynthesis

Motivation and Aim: The orthology relationship is not yet decisively formalized, and some of its important features may depend on the taxonomic context. Gene *moeB* is tackled as an example, and is itself an interesting object of research. It encodes an E1-like family enzyme involved in molybdopterin and thiamine biosynthesis. This family includes proteins that catalyze ATP adenylation of the C-terminal glycine carboxyl group in sulfur carrier proteins, e.g., MoaD or ThiS. Gene *moeB* is present in plastids of all sequenced red algae (Rhodophyta), with the exception of *Cyanidioschyzon merolae*.

Methods and Algorithms: All plastid proteins of the red algae available in GenBank, NCBI, were considered for the study. Analyses were conducted with original algorithms.

Results: We report a first complete database of plastid protein families from the red algae. The families contain proteins with maximal sequence similarity and minimal paralogous content, and were built based on an original definition of paralogy. The database contains 6005 protein entries, 495 families, 286 non-singletons (from which 235 are paralog-free, and other 51 contain at maximum two proteins per species). We will report results of a systematic comparison of the database with biological data and conclusions drawn from analyzing the database. Gene *moeB* exemplifies the results. Candidate bacterial-type promoters in 5'-leader regions of *moeB* distinctly differ from the template. In all sequenced species, except for *Cyanidium caldarium*, these regions contain a conserved 12 nt-long motif. Its high conserved part has the consensus TAGAT. In *Gracilaria tenuistipitata* the motif adjoins the -35 promoter box; in *Grateloupia taiwanensis* it coincides with the -35 box; in *Porphyridium purpureum* and *Chondrus crispus* the motif is distanced from the -35 box; in *Calliarthron tuberculosum*, *Porphyra purpurea*, *Pyropia haitanensis*, and *P. yezoensis* the -35 box is not determined, although the AT-rich region found downstream may represent a functioning -10 promoter box.

Conclusion: A notable distinction of the identified promoters from the consensus and presence of a closely located conserved motif suggest this motif to be a putative binding site of a transcription factor activating transcription of *moeB*. In most species, presence of an actively transcribed tRNA gene on the opposite strand precludes *moeB* transcription from a distantly located promoter. Lack of the motif and the bacterial-type promoter upstream *moeB* in *Cyanidium caldarium* suggests a different regulation mechanism for this gene, which comes in agreement with its loss in the close species *Cyanidioschyzon merolae*.

Availability: The database of plastid protein families from the red algae, http://lab6.iitp.ru/ppc/redline45

Acknowledgements: The study was supported by RFBR, research grant 13-04-40196-H.

Abstracts received after the deadline

PROBABILISTIC FORMAL CONCEPTS WITH NEGATION

Martynovich V.V. [1]*, Vityaev E.E. [2]
[1] *Novosibirsk State University, Novosibirsk, Russia;*
[2] *Sobolev Institute of Mathematics SB RAS, Novosibirsk, Russia*
* *Corresponding author*

Key words: *formal concepts, notions, data mining, association rules, classification*

Motivation and aim: The paper tends to generalize somewhat famous approach brought by Formal Concept Analysis [1, 2]. The main improvement related with including the negations into concept intent construction. Also statistical ambiguity problem should have been eliminaed by consistency and soundness results.

Methods and algorithms: The long ways starts from classic Formal Concept Analysis structures and tools [2]. Previous works [3] stated the definition of formal concept as fixed points for prediction operator. So the introduction of probability as in [5] on formal system leads to key notion of probabilistic concept. Using the ideas of semantic probabilistic inference [6], we propose some results about such concepts.

Results: The formal concepts with negations were introduced, prediction operator properties has been researched [6, 7], and fixed point consistency and soundness theorems has been proven, like in [8]. Additionally, the inconsistency case has been studied and described. Finally, those lead to algorithm, which yields a list of all probabilistic concepts.

Conclusion: Algorithm in paper could be significantly improved. Broad theoretical field of probabilistic formal context is a subject of subsequent researches. Presented method serves more attention and will be developed into independent Data Mining methods.

Availability: Full text is not available online right now, but may be provided by personal request via emailing.

References:

1. Ganter, B.: Formal Concept Analysis: Methods, and Applications in Computer Science. TU Dresden (2003)
2. Ganter, B., Wille, R.: Formal concept analysis. Mathematical Foundations. Berlin-Heidelberg-New York, Springer (1999)
3. Vityaev, E.E., Demin, A.V., Ponomaryov, D. K.: Probabilistic Generalization of Formal Concepts. Programming and Computer Software. 38(5), 219-230 (2012)
4. Missaoui, R., Kwuida, L.: Implications in Triadic Formal Contexts. In: 9th International Conference, ICFCA 2011. Nicosia, Cyprus, Springer (2011)
5. Vityaev, E., Kovalerchuk, B.: Empirical Theories Discovery based on the Measurement Theory. Mind and Machine 14(4), 551-573 (2004)
6. Vityaev, E.E.: Knowledge discovery. Computational cognition. Cognitive process models. Novosibirsk State University Press, Novosibirsk (2006) (in Russian)
7. Vityaev, E., Smerdov, S.: On the Problem of Prediction In: KONT/KPP 2007, LNAI 6581, pp. 280-296, Springer, Heidelberg, (2011)
8. Speransky, S.O.: Logic of probability and probability logic. Novosibirsk State University, Novosibirsk, PhD thesis. (2013) (in Russian)
9. Vityaev, E.E., Neupokoev, N.V.: Formal model of perception and pattern as fix point of anticipations. In: Approaches to the thinking modeling, pp.155-172, Moscow, URSS Editorial (2014)

MODELLING OF THE HEPATITIS C VIRUS LIFE CYCLE

Binder M.
German Cancer Research Center, Division Virus-Associated Carcinogenesis (F170), Im Neuenheimer Feld 280, D-69120 Heidelberg, Germany

Hepatitis C is a severe disease and a prime cause for liver transplantation. Up to 3% of the world's population are chronically infected with its causative agent, the Hepatitis C virus (HCV). This capacity to establish long (decades) lasting persistent infection sets HCV apart from other plus-strand RNA viruses that typically cause acute, self-limiting infections. A prerequisite for its capacity to persist is HCV's complex and tightly regulated intracellular replication strategy. In order to understand the molecular processes governing HCV RNA replication, using a combination of biological experiments and mathematical modeling, we developed the first detailed mathematical model of the initial dynamic phase of intracellular HCV RNA replication. For this purpose, we quantitatively measured viral RNA and protein translation upon synchronous delivery of viral genomes into host cells, and thoroughly validated the model using additional, independent experiments. Using this model to study HCV's replication strategy, we recognized diverse but crucial roles for the membraneous replication compartment of HCV in regulating replication by balancing translation versus replication and thus effectively limiting RNA amplification. We further predict the existence of an essential limiting host factor (or function) required for establishing active RNA replication and thereby determining cellular permissiveness for HCV. Our model also proved valuable to understand and predict the effects of pharmacological inhibitors of HCV and might be a solid basis for the development of similar models for other plus-strand RNA viruses. In currently ongoing work, we extend the model by further cellular processes, most importantly the innate antiviral response. This will eventually allow us to study- and ultimately interfere with- the intricate balance between viral exploitation of host cellular resources and intrinsic defense mechanisms of the cell, which is the basis of virus persistence.

INVASION, ADAPTATION AND EVOLUTION: WHEN ALL OUT OF SYNC

Suslov V.V.
Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
e-mail: valya@bionet.nsc.ru

Key words: Haldane's dilemma, invasion, evolution, preventive adaptations.

Motivation. Disagreements between temps of invasion, adaptation and evolution are a quick transition from slow to rapid invasion in new econiche/biotope and *vice versa*[1] (i), the relative irrelevance of a successful invasion of a new econiche to preadaptation to that econiche[2] (ii) and discordance between evolutionary and adaptation rates of small populations (SP) during invasion[2] (iii) [1]. *The Counterbalance of vectors of directional selection rule (CVDS-rule) explains i, but not ii, iii because applicable only to large populations*[3] [2]. The higher the number of genes under selection, the higher the death toll over a single generation (Haldane's dilemma)[4]. The reproductive reserve, which is, *ceteris paribus*, the smaller, the smaller the population size, should compensate for the deaths that have occurred *for all* reasons. In result, any increase in mutability, norm of reaction, or accidental death, including death due to invasion of the novelty or *even approaching the pessimal margin of the home econiche* are dangerous for SPs.

Results. Conception of "preventive adaptations" (PA) are formalized. PA not to be confused with preadaptations: PA do not counteract pessimal factors, but prevent any contact with them. Therefore, PAs are easily produced (important for SP!) but hardly eliminated by selection eventually become more sustainable than the function they protect[5] (see thesis II) and formed stasis in monomorphic population.

Acknowledgements: Budget Project VI.61.1.2., IP Presidium RAS No.28 «BOE».

References:

1. Suslov V.V. // XXVIII Liubischev Scientific Conference. Ulyanovsk. 2014. P. 118-126.
2. Severtsov A.S. // Zh. Obshch. Biol. 2012. 73, (5) P. 323-333.

[1] A prominent example is the rapidly (over ~5-30 years) mastering of perching on trees by urban pigeons, who had for centuries been preserving the inability to sit on tree branches as their wild rock dove ancestors.

[2] Isolating factors in Lake Baikal are a low temperature and low salinity. Although many of the endemic species spawn in the inflowing rivers (that is, *they are preadapted* to invasion and these preadaptations are tested i*n each generation*), invasions from Lake Baikal are rare. Per contra, invaders of any river origin – Russian sturgeon, pike, Amur catfish – have yielded viable morphs although they are not comfortable with baikalian water. By comparison, urban branch-sitting pigeons have not even formed morphs according to CVDS-rule though counterbalance of a light and a dark morphs is important in colonizing new biotopes.

[3] According to CVDS-rule stasis is persisted due to that selection by the pessimal factor is interfered by selection by another one. Consequently, any population is polymorphic i.e. consist of subpopulations adapted to the each pessimal factor and hybrids. Shift of the balance in selection vectors leads to a shift of balance of pure lines and hybrids [2]. But selection in many vectors is contraindicated for SP by Haldane's dilemma.

[4] Genome projects and the ENCODE project have demonstrated that populations are rich in SNPs and that pseudogenes and noncoding DNA can be synthesized. For most of its part, this variability is neutral, but the part that is not neutral jeopardizes the adaptations that be, and for that reason the first priority is to protect them rather than to search for rare useful alleles.

[5] Aphids bred using a novel food in G.Ch.Shaposhnikov's experiments established their food preference *after* survival on it: given a chance to invade, individuals with (pre)adaptations would have stopped invasion and would therefore have dropped the selection they benefit from.

ONTOLOGY OF PREVENTIVE ADAPTATIONS

Suslov V.V.
Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
e-mail: valya@bionet.nsc.ru

Key words: Haldane's dilemma, invasion, evolution, preventive adaptation, endemism

Motivation. According to above-noted Haldane's dilemma, the best small populations (SP) can do is get specialized, which leads to a *catastrophe over endemism*: from *moving back deeper into* predictable/comfortable areas of the econiche to settling down in those areas via "ontogenesis stabilization"[1] and\or via preventive adaptation (PA). In result, SPs must loss both opportunities and resources (due to areal collapse) for any invasion. It is a direct path to extinction. But SP can lead invasion in new econiche [1]. How is possible?

Results. PAs were classified into 7 types [1]: cyclic (*a*), cyclic on depletion (*b*), induced (*c*), inducible (*d*), nonspecific (*e*), constitutive (*f*) and combined. Type *a* PAs are exemplified by the gene network (GN) for circadian rhythms. Type *b* PAs are exemplified by GNs ranging from those for appetite to those for DNA repair, which act by synthesizing a pool of expendable by-products during the execution of the main function of the GN. Depletion of the pool reactivates the GN. A large number of neuromediators, messengers and other signaling molecules derive from the by-products of cellular metabolism. Type *c* PAs are represented by instincts (induced only by special realizer during special time[2]); type *d* PAs, by conditioned reflexes and the derivative cortical functions; type *e* PAs, by non-specific adaptive syndromes (pain as non-terminalized, and stress as terminalized). Type *f* PAs can be formed by selection from any trait[3]. Emergency of any type PA *efficiently promotes the increase of knowledge*[4], which can be attained either extensively (by improving outlook, which is knowing a bit about many things) or intensively (by improving pedantry, which is knowing many things about a bit). So, out of the 7 PA types, only *d and e enable increase outlook, while the most easily developed and commonest type f does not lead to an increase outlook or pedantry.*

Acknowledgements: Budget Project VI.61.1.2., IP Presidium RAS No.28 «BOE».

References:

1. Suslov V.V. // XXVIII Liubischev Scientific Conference. Ulyanovsk. 2014. P. 118-126.
2. Zhdanov A.A. Autonomous artificial intellect. M., Binom. 2009. 359 p. (in Russ.)

[1] This way abides by G.F.Gause's rule (interspecies relationships), A.S.Severtsov's CVDS-rule (see thes. I - intraspecies relationships) and A.G.Kreslavsky's rule of adaptive subniches (counteracting mostly abiotic factors): all rules relate to division of the *home niche, but not invasion of a new one* [1].

[2] The Western Capercaillie is a typical urbophobic species. Unfamiliar signals increase its anxiety and encourage this bird to retreat deeper into the biotope. ♀ pass their anxiety on to their young by singing before hatching. Listening to the songs vocalized by an unanxious ♀ reduced anxiety in their offsprings. A similar effect was produced by handling on pregnant rats or D.K.Belyaev's domestic foxes.

[3] Cliff swallows have had their wings shortened over 30 years as these birds have been hunting near highways: collisions with cars enabled selection for maneuverability in prejudice to all the other flying characteristics (in bush-loving passerines, the disadvantage of the analogous adaptation is compensated for by an energy-consuming style of flight). The onslaught of asphalt has resulted in decreased fertility and dissemination-related adaptations in urban populations of plants. Invasion by Asian ladybugs was withheld by inbred depression, which was later eliminated by wide breeding them as fighters against aphids [1].

[4] If the criterion for increase of knowledge is increase in predictability [2].

BALDWIN EFFECT AS EVOLUTION SILENCER

Suslov V.V.
Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
e-mail: valya@bionet.nsc.ru

Key words: Haldane's dilemma, invasion, evolution, preventive adaptation (PA), positive and negative Baldwin effect (BE)

Motivation. Traditionally attempts at resolving disagreements between rates of invasion, adaptation and evolution rely on optimization-based balance scenarios (OBS)[1] [1,2], and the BE [2,3]. Not denying success it should be stated that authors traditionally ignore Haldane's dilemma, small population (SP), PA any types[2] (see thesis Ontology of PA).

Results. The BE realized if the behavior and genetic adaptations are phenotypically similar or have common genetic foundation. Under this condition small and rare mutation must be phenotypically reinforced by modification and vice versa and the evolution/invasion rates must be accelerated [2,3]. However, if any PA would be one of the behavior or genetic adaptations, these rates be weakened (negative BE of the 1st kind). OBS demands the *preexisting* receptors and *large amounts of time* to enable *repeated* exploratory motives (EMs)[3]. Otherwise OBS dangerous. Therefore, in a SP any OBS will be blocked by selection favoring any PA[4] and the BE will only accelerate the blockade[5]. On the other hand, PA any types may be conditioning by behavior, organismal energy budget, etc[6]. If PA has weakened under the conditioning by a serial success of repeated EMs, the evolution/invasion rates in the uppish subpopulation must be accelerated but only with accidental death (negative BE of the 2nd kind, dangerous for SP). Finally OBS and the BE usage by a SP is problematic.

Acknowledgements: Budget Project VI.61.1.2., IP Presidium RAS No.28 «BOE».

References:

1. Zhdanov A.A. Autonomous artificial intellect. M., Binom. 2009. 359 p. (in Russ.)
2. Ridley M. The Red Queen. NY, Harper Perennial. 2003. 416 p.
3. Red'ko V.G. (2013) // Neyroinformatika-2013. Pt 2. M., MIFI. P. 257-266. (in Russ.)
4. Suslov V.V. // XXVIII Liubischev Scientific Conference. Ulyanovsk. 2014. P. 118-126.

[1] Under OBSs, selection favors the search for the novelty despite accidental death increase, because a cumulative acquisition of data should make the environment predictable [1,2].

[2] Under discuss only large populations with *preexisting* pool of receptors *suitable* for explored environment [1,2]. Their ability to invade in novelty is postulated *a priori* in models [1,3], or is strictly specified in experiments (as G.Ch.Shaposhnikov, see thes.I). Otherwise, selection favoring rare behavior (for example, animal taming) should be set for reception and pool of *preexisting* PA. Ignoring these condition leads to an inability to produce the response to selection at a population (even if the variance is high) or an individual (even if preadaptations are there) level [4].

[3] That is, OBS is good when invasion relates not to a new, but a phylogenetically familiar and comfortable econiche. Land bird migrations across the ocean, for example Cattle Egret to South America, just cannot have occurred under a series of repeated EMs [4].

[4] Especially *f*, see example with Cliff swallows in thesis Ontology of PA.

[5] For example, the brown bear is kept away from the brachiator's econiche by fear. Another example with Western Capercaillie see thes.II. Plants (passive disseminators) invade from recently island biotopes (for real island, forest island in steppe or park in city) more efficiently than animals with their neophobia and fear [4].

[6] Some types (*b, d,* stress as *e*) amenable conditioning better than others (*a, c, f,* pain as *e*).

INDIRECT SELECTION AS INVASION SILENCER

Suslov V.V.
Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
e-mail: valya@bionet.nsc.ru

Key words: invasion, evolution, indirect selection (IS), preventive adaptation (PA)

Motivation. Another well-known attempt to resolve above-noted disagreements between invasion, adaptation and evolution is indirect selection (IS) when one adaptation increases the probability of another[1], with the formation of multi-step cycles including the coming into phylogenetically unfamiliar or poorly efferent environment[2]. IS scenario has been proposed as the solution to the problem of the quickly invasion and adaptive evolution into lands landscape by chordates[3] [2] and arthropods [3] (Fig.1, cycle 1.-2.-3.-4.-5.).

Results. IS scenario may facilitate adaptive evolution *after* the invasion in new econiche but not initiate the invasion due to *bypasses* (Fig. 1, other cycles)*:* it explains well replacements of organs in a new and/or a phylogenetically familiar environment (such replacements are nothing else than pre- and post-adaptations), but not the replacement of environments. Really, modern fish with euadaptations to air-breathing are inclined to rely on the "whale strategy": feeding in an oxygen-poor *hydro*biotope↔resting in an oxygen-rich *hydro*biotope (the *water* surface for air-breathers and the for example surface *water* for those breathing with gills[4]). In natural conditions, during a drought Australian lungfish does not come out of water; other dipnoans outside water are in anabiosis, although in captivity, i.e., in optimal conditions they can travel on land [4]. Preexisting PAs started by pessimization and blocked any invasion that may be happen without it.

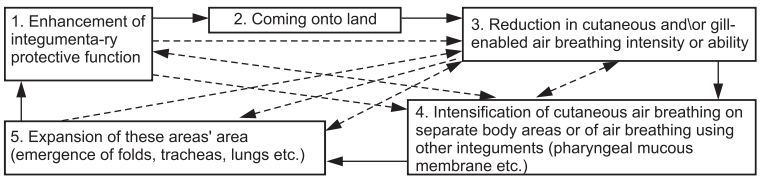


Fig. 1. Indirect selection and invasion of dry land (⟶➤). Cycle **1.→2.→3.→4.→5.** [2, 3]) is diverted (⟶➤) to bypass vertex **2.** in modern air-breathing fish.

Acknowledgements: Budget Project VI.61.1.2., IP Presidium RAS No.28 «BOE».

References:

1. Afonnikov D.A., Koichanov N.A. (2001) // Dokl. Biochem. Biophys. 380. P. 352-355.
2. Simpson G.G. Tempo and Mode in Evolution. M., IL. 1948. 369 p. (in Russ.)
3. Gilyarov M.S. Common Patterns in Dry-Land Adaptations by Arthropods. M., Nauka. 1970. 276 p. (in Russ.)
4. Suslov V.V. (2013) // Paleontology Journal (Mosk). 47, (9). P. 1048-1055.

[1] Like scenario of the coadaptive substitutions in molecular evolution [1] but this scenario does not include any PAs.

[2] Environment for perception of which no suitable receptors or the resolving capacity of the preexisting receptors/acceptor of the result is insufficient or when they are slow to activate.

[3] Well-known scenario of the additive accumulation of minor (pre)adaptations in the ecotone is a serial cycles 1-2.-3.-4.-5. (see Fig.) with more and more dry biotope on the vertex 2. But irrespective of the type of biotopes it is primarily necessary for potential invaders to approach to the biotopes [4].

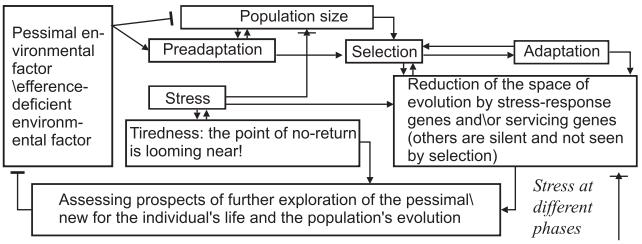
[4] Cases of the "whale strategy" have also been spotted in aerated water containing matter that is bad for gills. For example, so are the waters in African thermal soda lakes or in Russia's rivers during the so-called "dark-water spring floods" [4].

INVASION, ADAPTATION AND EVOLUTION: HOW TO RECONCILE

Suslov V.V.
Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
e-mail: valya@bionet.nsc.ru

Key words: invasion, evolution, preventive adaptation (PA), small population (SP)

Motivation. Finally, any trends of evolution[1] without population growth are dangerous for SPs and specialization is the lesser evil only. Only preexisting adaptive pool of receptors, modifications, etc. are predetermines Baldwin effect or indirect selection as adaptive. Yet, the Osborn effect[2] is exist, SPs are leads an invasion [3].



Results. Previously proposed *par force* evolution scenario, the organisms adapts to its non-specific terminalized PA, i.e. stress-reaction (by prolongation but not optimization[3] of its cross-resistance phase - CR, reduction of distress, and acceleration of the alarm phase) rather than environments. Stress automatically reduces the diversity of expressed genes (including expression of PA non-*f* types) to stress-response genes and genes of the main servicing infrastructures (often they are one and the same). These genes are low in number; however, their fitness has many times been tested during the phylogenesis because of their non-specificity (hence the fast-paced evolution in SPs and parallelisms as noted by Osborn). Reduction of the space of evolution increases evolutionary rates, decreases the cost of selection, *temporarily*[4] allows overcoming disadvantage of the f type PA in novelty. Using stress an individual can *in situ,* but not *a posteriori* as in Darwinian selection, assesses how perilous the novelty is based on the amount of tiredness[5], even if that individual has no receptors for the new pessimal factors (see fig.) [3].

Acknowledgements: Budget Project VI.61.1.2., IP Presidium RAS No.28 «BOE».

References:

1. Koonin E.V. The Logic of Chance. FT Press Science. 2011. 528 p.
2. Suslov V.V., Kolchanov N.A. (2009) // Vestnik VOGiS. 13, (2). P. 410-439. (in Russ.)
3. Suslov V.V. (2013) // Paleontology Journal (Mosk). 47, (9). P. 1048-1055.

[1] Including neutral evolution or genetic drift because PAs (*b* and *f* types especially) may provoke weakening of the selection or areal collapse, but simultaneously forbid the testing of any neutral increase of the variability in new environments. Only under the possibility of the testing these neutral evolution or drift can help to find new peaks of the Wright's adaptive landscape [1]. PAs (except types e) can neutralize non-neutral mutations as negative feedbacks or slow reactions in homeostatic gene networks [2] but in contrast of ones by the preventing any contact with pessimal factors not to phenotype stabilization. Consequently, after formation of PAs its importance can reinforced by neutral evolution and vice versa.

[2] For the H.F.Osborn who the first notes regularities of these invasions [3].

[3] This selection leads to a catastrophe over endemism due to favors the optimization of any phase of stress-reaction in prejudice of the others or reduces non-specificity by converting stress into another PA type.

[4] During CR animal can do *brief* excursions to various pessimal habitats. This is reduced negative Osborn effect (identified during acclimatization) when SP was die because animals melted away.

[5] Tiredness results from an interference of resources, rather than from their deficiency due to develops long before the deficiency.

SIZE MATTERS: MATHEMATICAL MODELING OF THE Type-1 INTERFERON INDUCTION

Kaderali L.
TU Dresden, Medical Faculty, Institute for Medical Informatics and Biometry, Dresden, Germany

Retinoic acid-inducible gene I (RIG-I) is a cytoplasmic innate immune receptor for viral RNA. RIG-I detects viral RNA and triggers an antiviral response by activating the type I interferon (IFN) pathway. While it is known that a 5′-triphosphate end, as well as double-strandedness of RNA constitute molecular patterns recognised by RIG-I, the underlying mechanism is less well established. Recent experiments show a RNA length-dependent activation of the IFN response, with long dsRNA evoking a larger immune response than short RNA at non-saturating concentrations. We developed a mathematical model of viral RNA recognition by RIG-I, as well as the activation of downstream players in the pathway. Our model is able to explain quantitative activation curves of RIG-I enzymatic activity, as well as the cellular response of IFN transcription depending on RNA length and concentration. Furthermore, it is consistent with additional data from structural microscopy and size exclusion experiments. Our model predicts an important role of cooperative binding in RNA recongnition, showing how RIG-I has evolved to sensitively detect cellular infections by RNA viruses such as influenza. The cooperative binding of RIG-I along the length of RNA molecules ensures a large immune response even at low copy numbers of viral RNA genomes.

REPEATED POSITIVE FIGHTING EXPERIENCE IN MALE MICE: THE TOOL FOR THE STUDY OF MOVEMENT DISORDERS

Kudryavtseva N.N.*[1], Smagin D.A.[1], Kovalenko I.L.[1], Vishnivetskaya G.B.[1], Galyamina A.G.[1], Enikolopov G.N.[2]
[1]*Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia*
[2] *NBIC, Moscow Institute of Physics and Technology, 123182, Moscow, Russia,*
e-mail: n.n.kudryavtseva@gmail.com
[*]*Corresponding author*

Key words: repeated aggression, brain dopaminergic systems, movement disorders

Repeated aggression is a frequent symptom of many psychiatric and neurological disorders. However, this phenomenon is poorly studied because of the lack of adequate models in animals. It has been shown that male mice with a long positive fighting history develop behavioral psychopathology, which includes abnormal aggression, strong hostility, pronounced anxiety, disturbances in social recognition and motivated behaviors, impairments of sociability etc. (Kudryavtseva, 2006). Behavioral observations suggest that positive fighting history provides a permanent reward to male mice, hence a tendency to repeat aggression acts. Male mice with prolonged aggressive experience demonstrate also movement disturbances: hyperactivity, attention/deficit behavior, motor dysfunctions and repetitive stereotypic behaviors, such as repulsive behaviors, jerks, rotations, jumping, head twitches etc. It has been shown that balance between the activities of the brain's neurotransmitter systems is disturbed in animals under repeated aggression due to a reduced activity of the serotonergic system and an enhanced activity of the dopaminergic systems in the brain. In the winners, expression of dopaminergic genes (*Th, Dat1,Snca*) is upregulated in the ventral tegmental area as compared with the controls. In the midbrain raphe nuclei the expression of serotonergic genes (*Tph2, Sert, Maoa, Htr1a*) is downregulated under repeated aggression in male mice. Our data support the idea that chronic activation of brain dopaminergic systems is supposedly responsible for increase of aggressiveness and disturbances in movement activities, which may be considered as signs of developing psychomotor or neurologic disorders such as, for example, Tourette's syndrome, autistic spectrum disorders or Huntington's disease.

Acknowledgements: The work is supported by Russian Scientific Foundation (14-15-00063).

References:

1. Kudryavtseva N.N. (2006) The psychopathology of repeated aggression: a
2. neurobiological aspect. In "Perspectives on the Psychology of Aggression". Ed. JP
3. Morgan. NY: Nova Science Publichers, Inc. Chapter 2. p.35-64.

GETTING SOPHISTICATED: NEW APPROACHES, TRENDS AND DEVELOPMENTS FOR DDPCR

Verner A.
Bio-Rad Laboratories, Moscow

In recent years droplet digital PCR (ddPCR) became very popular method for quantitative DNA analysis. The areas where ddPCR technology can utilize its unique characteristics can was defined. Superior ability of ddPCR in detection of rare events and copy number variations of genes propelled researchers from oncology field to begin developing diagnostics of next-level of sensitivity and accuracy, including liquid biopsy. High versatility of ddPCR attracted scientists involved in projects of various interests, including GMO detection, environmental studies, and infection detection developments. The method can be used for quantitative and qualitative library analysis for NGS. Now, researchers learn how to ask questions only digital PCR can answer, and currently number of digital PCR applications growth steadily.

COMPUTER STUDY OF GENE EXPRESSION RELATED TO AGGRESSIVE AND TOLERANT BEHAVIORS ON LABORATORY ANIMALS

Orlov Y.L.[1,2]*, Spitsina A.M.[2], Medvedeva I.V.[1], Bragin A.O.[1], Anikeev A.V.[2], Galyamina A.G.[1], Kozhemyakina R.V., Safronova N.S., Kovalenko I.L., Konoshenko M.I.[1], Moreva T.A.[1], Kudryavtseva N.N.[1], Markel A.L.[1,2]

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

[2]*Novosibirsk State University*

e-mail: orlov@bionet.nsc.ru

[*]*Corresponding author*

Key words: aggressive behavior, anxiety, mouse models, gene expression, transcriptome sequencing, microarrays

Motivation and Aim: Nature of human aggression is complex in its manifestation and could be studied with any limitations on animal models. Excessive aggression is the major problems of the human society and health, since repeated aggression is observed in the many psychoemotional and neurological disorders such as bipolar (manic-depressive) psychosis, epilepsy, schizophrenia, Tourette's syndrome, autism. Basic studies have shown that the frequency and severity of aggression depends on the hereditary predispositions, previous experience of aggressive behavior and social context. In literature, the main conclusion underlined a multi-loci determination of aggressive behavior. There are no major genes reported. We started studying of the molecular genetic mechanisms of enhanced aggressiveness in comparison with tolerant behavior using two unique experimental models developed at ICG SB RAS: gray rats (*Rattus norvegicus*) selected for aggressive and tolerant behavior and mice reared for enhanced intermale aggressiveness.

Methods and Algorithms: Microarrays and next generation sequencing technology (RNA-Seq) will be used to study the transcriptome in different brain regions from aggressive and tame animals. The following brain areas involved in regulation of aggressive and tolerant behavior will be examined: prefrontal cortex, ventral tegmental area containing the dopaminergic and opioidergic neurons and responsible for the positive reinforcement, midbrain raphe nuclei, and the amygdala regulating emotions. We started analysis of gene expression at these brain areas using BioGPS database (biogps.org) [1].

Results: Computer program for analyzing of gene expression correlations is written. We analyzed gene expression in human brain [1] and selected specific lists of genes over- and under-expressed in brain tissues by BioGPS. We used genome alignment data to reveal gene regions conserved in human and rat and in human and mouse for the selected genes. Application of the whole transcriptome sequencing analysis for these brain regions of experimental animals is intended to assess the significance of genetic and environmental factors in determination of enhanced aggression.

Acknowledgements: The work is supported by RSF (14-14-00269).

References:

1. I.V. Medvedeva et al. (2013) Computer analysis of data on gene expression in brain cells obtained by microarray tests and high-throughput sequencing // Vavilov Journal of Genetics and Breeding, V.17, N. 4/1, p. 629-638

ION TORRENT™ PLATFORM IN 2014: TECHNOLOGY AND APPLICATIONS

Natalin P.B.*
Genetic Analysis Team Leader, Genetic Medical & Applied Sciences, Life Sciences Solutions, Thermo Fisher Scientific, Moscow, Russia
e-mail: pavel.natalin@thermofisher.com
**Corresponding author*

Key words: *NGS, semiconductor sequencing, exome sequencing, RNA-Seq, AmpliSeq™*

Motivation and Aim: The results of the latest development of Ion Torrent™ platform for semiconductor sequencing are described. Main application of Ion PGM™ (microbial sequencing and targeted re-sequencing) and Ion Proton™ (exome and transcriptome sequencing) are discussed.

Methods and Algorithms: Electronic engineering, *in vitro* enzyme evolution, software engineering.

Results: Technological improvements of the platform include a new template preparation system Ion Chef™ and reagents, new high-accurate Hi-Q™ sequencing chemistry, Ion PII Chip data and availability, AmpliSeq™ DNA and RNA panels, Ion Reporter Server and enhanced data analysis software capabilities.

Conclusion: Innovation can be unpredictable, but it is inevitable. Ion Torrent™ platform undergoes constant development which drives the adoption of NGS for fundamental research, translational medicine and clinical studies.

Availability: Ion Torrent™ data analysis software is freely available over the Internet on *www.ampliseq.com, https://ionreporter.lifetechnologies.com*

References

1. The results presented are internal Research & Development results of Thermo Fisher Scientific.

CHARACTERISTICS OF TYPE 1 DIABETES SUSCEPTIBILITY REGIONS

te Boekhorst R.*, Beka S., Abnizova I.
Bio-computation Group, School of Computer Science, University of Hertfordshire, UK
e-mail: r.teboekhorst@herts.ac.uk
**Corresponding author*

Key words: *medical genetics*, *disease – associated SPNs, intronic and regulatory DNA*

Motivation and Aim: Diseases like diabetes, neurodegenerative diseases and cancers are called "complex diseases" because their instigation includes a combination of multiple genetic, environmental and life style factors. To study the genetic factors, scientists have traditionally focused on finding single point mutations (SNPs) in the genes of those suffering from the disease. However, the role of disturbed gene regulation rather than disrupted protein coding is increasingly recognized. Yet, even if regulatory aspects are acknowledged, just the identification of associated SNPs, whether they occur in genes or in regulatory modules, by means of a GWAS is just one step in unravelling the genomic aetiology of complex diseases.

Methods and Algorithms: To build a broader picture, we suggest characterising areas that are known to confer susceptibility to a particular complex disease by features that capture their distinctive genomic entirety on a higher level of organisation than mere location. Using Type 1 Diabetes (T1D) as an example, we analysed a set of genomic variables in order to typify the susceptibility regions connected with this disease. The aim was to find out if particular regions differ strikingly in genomic content from others and if so what structural properties are responsible for this classification and whether it is i) associated with a distinction in reputed functional features and ii) reflected in the concomitance of other autoimmune diseases.

Results: The analysis revealed two main clusters of susceptibility regions and hence suggests the existence of two types of genomic areas that are differently involved in the occurrence of T1D and associated auto-immune diseases. The first cluster consists of regions that contain large sequences of intronic and regulatory DNA. Especially these susceptibility regions are also loci for many other autoimmune diseases. The second cluster, which includes the Human Leukocyte Antigen locus, is made up of two sub-clusters. The first comprises short, gene dense regions with high SNP counts. These regions are also rich in SNPs occurring in experimentally verified transcription factor binding sites and – like cluster 1 - are loci for many co-occurring auto-immune diseases. The second sub-cluster cluster had no particular outstanding attributes and only a small number of its regions are loci for any of the 17 co-occurring auto-immune diseases.

Reference

1. O.S. Burren et al. (2011) T1DBase: update 2011, organization and presentation of large-scale data sets for type 1 diabetes research. Nucleic Acids Res. 39:D997-1001

INFLUENCE OF CONSUMABLES ON QUALITY AND PRECISION OF EXPERIMENTS

Dzhenin S.V.,
Limited Liability Company Eppendorf, Russia

Pipette tips, microcentrifuge tubes and plates are always used in all laboratories. The researchers often do not consider how ordinary consumables influence biochemical reactions, samples during storage and ultimately results of research. Leaching of different agents from plastic, binding samples with the walls of tubes, inaccurate pipetting complex samples - all this has an impact on the results of research. Eppendorf consumables will not only save the sample but also provide integrity of test and improve accuracy and reproducibility of results. Besides it will save money and time.

Author index

L

M

N

O

P

R

S

T

V

W

Y

Z

Компания Хеликон – крупнейший поставщик в России самого современного оборудования и реактивов для широкого спектра научных исследований, молекулярной диагностики, криминалистики. Клиентами Компании Хеликон являются исследователи в области молекулярной и клеточной биологии, протеомике, молекулярной диагностике, медико-генетической экспертизе и идентификации личности, а также ветеринарии и селекции в сельском хозяйстве.

Мы являемся дистрибьютором ведущих мировых брендов:

Life Technologies,
Thermo Fisher Scientific,
BioRad,
Eppendorf,
Affymetrix,
Fluidigm,
SSI и др.

Почему клиенты работают с нами:

- Мы динамичная, быстро развивающаяся компания.
- Мы являемся не просто дистрибьютором ведущих мировых производителей, но и их методическим и сервисным центром в России.
- Мы всегда на шаг впереди использования и внедрения новых методов и технологий в молекулярной биологии.
- Мы тесно работаем с ведущими биотехнологическими компаниями и известными институтами.
- Мы постоянно проводим практические школы по NGS секвенированию, цитометрии, микроскопии.
- Мы работаем по всей России, имея представительства в 7 регионах.
- Мы обеспечиваем обучение персонала и постоянную методическую поддержку, технический сервис, своевременную доставку оборудования и расходных материалов.

119991 Москва, Ленинские горы, МГУ, д. 1, строение 40,
НИИ Физико-химической биологии им. А.Н.Белозерского, Лабораторный корпус «А»
Телефоны: +7 (499) 705-50-50, +7 (499) 769-51-60 Факс: +7 (495) 930-00-84
mail@helicon.ru, helicon.ru

Представительство в Сибирском регионе:
г. Новосибирск, ул. Инженерная, д. 28
+7 (383) 207-84-85, novosibirsk@helicon.ru

Представительство в Северо-Западном регионе: г. Санкт-Петербург, ул. Ушинского, д. 5, корп. 2, лит. А, пом. 1-Н.
+7 (812) 244-85-52, spb@helicon.ru

Представительство в Южном регионе:
г. Ростов-на-Дону,

Представительство в Приволжском регионе:
г. Казань, ул. Петербургская,

Компания Genotek оказывает услуги по высокопроизводительному секвенированию и генотипированию различных организмов с 2010 г. и является ведущим поставщиком на рынке генетических услуг.

Основателями компании являются выпускники биологического и механико-математического факультетов МГУ им. М.В. Ломоносова. В Genotek работают выпускники ведущих вузов России: МГУ, МИФИ, РГМУ им. Пирогова, МГТУ им. Н. Э. Баумана и др. Многие из нас имеют опыт работы в ведущих международных лабораториях.

Наши клиенты - ведущие научно-иследовательские институты и фармкомпании России и стран СНГ, такие как Институт общей генетики им. Вавилова, Институт цитологии и генетики РАН, Московская медицинская академия им. И. М. Сеченова и др.

Благодаря доступности передовых платформ для секвенирования, Genotek предлагает широкий спектр возможностей для Ваших исследований. А значительный объем заказов, получаемых компанией, позволяет нам оптимально использовать производственные мощности оборудования, чтобы всегда предлагать Вам самые низкие цены.

Подробнее об услугах: www.genotek.ru/R&D

Phone: +7(495) 767 85 95 или +7 (985) 411 64 07

Email: info@genotek.ru или ok@genotek.ru

ООО «Био-Рад Лаборатории»
117105, Москва, Варшавское шоссе, д.9 стр. 1Б,
Офисный комплекс Loft - квартал "ДМ - 1867"
ТЕЛ: (495) 721-1404
ФАКС: (495) 721-1412
post@bio-rad.com
www.bio-rad.com

Компания **Bio-Rad Laboratories, Inc USA** (Био-Рад, США) является одним из мировых лидеров производства оборудования и реагентов для научных исследований. В рамках взаимодействия с научными, медицинскими, биотехнологическими и образовательными организациями Био-Рад предлагает современные технологии, оборудование и реагенты.

Геномные технологии *(генная экспрессия и генная модуляция)*

- **Амплификация** *(уникальный спектр приборов)*
- **Цифровой капельный ПЦР третьего поколения**
- **Гельэлектрофорез** *(горизонтальный и вертикальный форматы)*
- **Experion™** *(автоматизированная система капиллярного электрофореза нуклеиновых кислот и белков)*
- **Системы визуализации** *(колориметрия, флюоресценция, хемилюминесценция, хемифлюоресценция, радиоизотопные метки)*
- **Перенос генов** *(электропорация, болистика, химическая трансфекция)*

Протеомные технологии *(структурная и функциональная протеомика)*

- **BioLogic DuoFlow™** *(модульная гибкая система для биохроматографии)*
- Широкий спектр колонок и носителей
- **NGC ™** *(автоматизированная хроматографическая система очистки рекомбинантных белков)*
- **Rotofor™** (аналитический и препаративный изоэлектрофокус)
- Оборудование для анализа и процессинга 2-D протеомных карт
- **Bio-Plex MagPix** *(мультиплексный количественный анализ биомолекул, панели для определения цитокинового профиля, белков сигнальной трансдукции, реагенты для создания собственных уникальных наборов)*
- **ProteOn XPR36™ Protein Interaction Array System** (*матричный интерактивный анализ биомолекулярных взаимодействий методом поверхностного плазмодинного резонанса*)
- **S3 сортер клеток**

Обзор инструментов
http://software.intel.com/en-us/intel-sdp-home/

Пакеты XE Suites

Intel® Parallel Studio XE
http://software.intel.com/en-us/articles/intel-parallel-studio-xe/

Intel® Cluster Studio XE
http://software.intel.com/en-us/articles/intel-cluster-studio-xe/

Статьи и материалы для самостоятельного изучения

Технические вебинары по программным инструментам
http://software.intel.com/en-us/articles/intel-software-tools-technical-webinar-series

Документация
http://software.intel.com/en-us/articles/intel-software-product-tutorials/#studioxe

Сертификационная программа
http://www.intuit.ru/academies/companies/972/info

Учебная лаборатория
http://software.intel.com/en-us/articles/intel-learning-lab/

Intel Premier Support
http://software.intel.com/en-us/articles/performance-tools-for-software-developers-intel-premier-support/

ООО «Рош Диагностика Рус»

Оборудование и реагенты для молекулярно-биологических, геномных и клеточных исследований

Компания предлагает оборудование и реагенты для лабораторий, работающих в области молекулярной и клеточной биологии, а также молекулярной диагностики:

- системы для секвенирования нового поколения 454
- системы для ПЦР в реальном времени LightCycler
- системы экстракции нуклеиновых кислот MagNAPure
- ДНК-чипы для обогащения образцов перед секвенированием нового поколения NimbleGen Sequence Capture
- клеточные анализаторы Innovatis
- высококачественные наборы реагентов

Компания предлагает покупателю не просто свою продукцию, а единую систему, включающую технический сервис, обучение персонала, постоянную методическую поддержку, своевременную доставку реактивов и расходных материалов.

ООО «Рош Диагностика Рус»
Россия, 115114, Москва
Летниковская ул., д.2/2
Бизнес-центр «Вивальди Плаза»
Тел.: +7-495-229-69-99
Факс: +7-495-229-62-64
ras.russia@roche.com
www.roche-applied-science.ru
www.roche-applied-science.com

Россия, 630132, Новосибирск
Ул.Красноярская, д.35, офис 103
Бизнес-центр «Гринвич»
Тел./Факс: +7-383-335-88-04/03
e-mail: olga.yatsenko@roche.com

Научное издание

Тезисы IX Международной конференции «Биоинформатика регуляции и структуры генома\ Системная биология»

на английском языке

Abstracts of the Ninth International Conference on Bioinformatics of Genome Regulation and Structure\ Systems Biology

Printed without editing

Дизайн и компьютерная верстка: А.В. Харкевич

Подписано к печати 30.05.2014 г.
Формат бумаги 70 × 108 1/16. Печ. л. 17,8. Уч.-изд. л. 19,7
Тираж 350. Заказ 120

Отпечатано в типографии ФГУП «Издательство СО РАН»
630090, Новосибирск, Морской пр., 2