RUSSIAN ACADEMY OF SCIENCES
SIBERIAN BRANCH

INSTITUTE OF CYTOLOGY AND GENETICS

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

Abstracts

BGRS\SB'12
Novosibirsk, Russia
June 25–29, 2012

Novosibirsk
2012

INTERNATIONAL PROGRAM COMMITTEE

Nikolay Kolchanov Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia (Chairman of the Conference)
Ralf Hofestaedt University of Bielefeld, Germany (Co-Chairman of the Conference)
Konstantin Skryabin "Bioengineering" Center, RAS, Moscow, Russia (Co-Chairman of the Conference)
Tamara Khlebodarova Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia (Academic Secretary)

Dmitry Afonnikov Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
Yuriy Aulchenko Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
Ming Chen Zhejiang University, Hangzhou, China
Roman Efremov Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry of RAS, Moscow, Russia
Frank Eisenhaber Bioinformatics Institute, Singapore
Fazel Famili University of Ottawa, IIT/ITI - National Research Council Canada, Ottawa, Canada
Vladimir Golubyatnikov Sobolev Institute of Mathematics, Novosibirsk, Russia
Igor Goryanin Biomedical Cluster Skolkovo, Russia
Vladimir Ilyin SINP MSU, Moscow, Russia
Vladimir Ivanisenko Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
Jaap Kaandorp University of Amsterdam, Netherlands
Lars Kaderali Dresden University, Germany
Olga Krebs Heidelberg Institute for Theoretical Studies, Germany
Alexey Kochetov Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
Vsevolod Makeev Vavilov Institute of General Genetics RAS, Moscow, Russia
Eric Mjolsness University of California, Irvine, USA
Mikhail Moshkin Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
Egor Prokhorchouk "Bioengineering" Center, RAS, Moscow, Russia
Valery Puzyrev Research Institute of Medical Genetics SB RAMS, Tomsk 634050, RUSSIA
Alexander Ratushny Institute for Systems Biology and Seattle Biomedical Research Institute, Seattle, WA, USA
Igor Rogozin National Center for Biotechnology Information, National Institutes of Health, Bethesda, USA
Nikolay Rubtsov Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
Andrey Rzhetsky University of Chicago, USA
Maria Samsonova St.Petersburg State Polytechnic University, St.Petersburg, Russia
Oleg Serov Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
Elena Schwartz Johns Hopkins University/Ariadne Diagnostics LLC, MD, USA
Victor Solovyev Department of Computer Science, Royal Holloway, University of London, UK
Evgenii Vityaev Sobolev Institute of Mathematics, Novosibirsk, Russia
Mikhail Voevoda Institute of Internal Medicine SB RAMS, Novosibirsk, Russia
Edgar Wingender University Medical Center Goettingen, Dept. of Bioinformatics, Goettingen, Germany
Limsoon Wong National University of Singapore, Singapore
Jaroslaw Zola Electrical & Computer Engineering, Iowa State University, Ames, IA, USA

LOCAL ORGANIZING COMMITTEE

Svetlana Zubova Institute of Cytology and Genetics, Novosibirsk, Russia (Chairperson)
Ilya Akberdin Institute of Cytology and Genetics, Novosibirsk, Russia
Yuriy Orlov Institute of Cytology and Genetics, Novosibirsk, Russia
Erlan Tokpanov Institute of Cytology and Genetics, Novosibirsk, Russia
Nadezhda Glebova Institute of Cytology and Genetics, Novosibirsk, Russia
Tatyana Karamysheva Institute of Cytology and Genetics, Novosibirsk, Russia
Andrey Kharkevich Institute of Cytology and Genetics, Novosibirsk, Russia
Galina Kiseleva Institute of Cytology and Genetics, Novosibirsk, Russia
Victoria Mironova Institute of Cytology and Genetics, Novosibirsk, Russia
Anna Onchukova Institute of Cytology and Genetics, Novosibirsk, Russia

ISBN 978-5-91291-011-1
© INSTITUTE OF CYTOLOGY AND GENETICS SB RAS, 2012

Organizers

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

Siberian Branch of the Russian Academy of Sciences

Russian Foundation for Basic Research

Program of SB RAS "Genomics, Proteomics, Bioinformatics"

The Vavilov Society of Geneticists and Breeders

Department of Systems Biology

Chair of Information Biology

PBSoft Ltd.

Sponsors

Russian Foundation for Basic Research

Skolkovo Innovation Centre

GOLD SPONSOR

Life Technologies

SILVER SPONSORS

Bruker Corporation

Agency "Khimexpert"

Hewlett Packard

"OPTEC", LLC

OOO «Miass Factory of Medical Equipment» & ZAO «Aceptic Medical Systems»

BRONZE SPONSORS

intel

Bio-Rad

Roshe diagnostics

Contents

SYSTEMATIC ERRORS AND BIASES IN ILLUMINA SEQUENCING

Abnizova I.I., Leonard S., Skelly T., Jackson D.
Wellcome Trust Sanger Institute, Hinxton, UK

Key words: *NGS Illumina sequencing, systematic errors, context-dependency*

Motivation and Aim: Sequencing of individual genomes and the determination of rare variants across populations are enabled by whole genome sequencing at low cost [1]. However, whole genome sequencing is accompanied by higher error rates [2,3] than capillary sequencing. Improved methods that accommodate these high error rates are needed in the calling of heterozygous sites from low coverage data. The design of effective statistical methods requires precise characterization of error in high-throughput sequence data.

Together with well known error tendencies in Illumina sequencing: phasing inaccuracy and cross-talk,- there is considerable amount of evidence about context and positional dependency for Illumina errors now. In addition to context dependency, there was found [4] strong variability of error profiles and rates between different Illumina versions (v4 vs v5), machines, runs and even the first and second reads in a pair for the same run and machine.

The error rate for base calling is commonly measured by confidence value (QV Phred [5]) of a base call. The process of generating QV for a base call is done by so called 'calibration' method. There are several well known calibration methods accepted in NGS community, the most established is PHRED method, which is implemented by Illumina's (Gerald and Bustard), and by GATK re-calibration.

*Methods and Result***.** In this work we introduce simple, fast PHRED-inspired method of calibration, PB-calibration, which takes care of mentioned above error dependencies and variability. We compare common recalibration methods, and show how run-specific PB-calibration helps to distinguish systematic errors from possible variants. The PB-calibration of a base call, combined with mismatch recalibration, can reliably identify around 30-40% of high quality mismatches, which cannot be marked with any other calibration methods, so far as we know. In general, PB-calibration may be applied to any NGS platform. It is currently reference-based, but can be easily adjusted to contig-scuffold-based (or any other) assembly method. PB-based error correction can be applied even without any available reference.

*Availability***.** PB-calibration is currently implemented in the production pipeline for spiked runs at Sanger Institute, and is publicly available from authors by request.

References

1. Nielsen R: Genomics: In search of rare human variants. Nature 2010,467(7319):1050-1051.
2. Hoff K: The effect of sequencing errors on metagenomic gene prediction. BMC Genomics 2009, 10:520
3. Dohm JC, Lottaz C, Borodina T, Himmelbauer H: (2008)Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Research, 36(16):e105.
4. McElroy et al. (2012) GemSIM: general, error-model based simulator of next-generation sequencing data, BMC Genomics, 13:74
5. Ewing, B. & Green, P.(1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. 8, 186–194.

ANALYSIS OF SEQUENCE FEATURES SPECIFYING THE ADHESION ABILITY OF INFLUENZA A VIRUS NEURAMINIDASE AND HEMAGGLUTININ

Afonnikov D.A.*[1, 2], Ivanisenko V.A.[1], Ignatieva E.V.[1], Medvedeva I.V.[1], Demenkov P.S.[1], Ivanisenko T.V. [1], Shah A.R.[3], Ramachandran S.[3]
[1] *Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia;*
[2] *Novosibirsk State University, Novosibirsk, Russia;*
[3] *CSIR-Institute of Genomics and Integrative Biology, Delhi, India*
e-mail: ada@bionet.nsc.ru
** Corresponding author*

Key words: *adhesion, sequence-activity relationship, molecular evolution, influenza A, neuraminidase, hemagglutinin*

Motivation and Aim: Flu epidemics caused by Influenza A is one of the major problems of health worldwide. Despite significant success in protection achieved due to vaccination, new mutant genotypes emerge sporadically and some of them Flu pandemics. Large scale sequencing of genomic sequences from virus isolates provides useful information about the variation and evolution of viral genomes. One possible approach is to investigate adhesive properties of viral proteins that aid host-pathogen interactions using bioinformatics methods. These investigations have potential towards developing strategies to weaken the ability of invasion of viral particles into host organism.

Methods and Algorithms: In this work we used SPAAN [1] to estimate adhesion indices (P_{ad}) for influenza A hemagglutinin and neuraminidase proteins in more than 8800 viral genotypes from human and four animal viruses sourced from The NCBI Influenza virus resource [2]. Sequence analysis was performed using WebProAnalyst[3].

Results: We observed clear differences in P_{ad} values between viral genomes of different subtypes/hosts/isolation times. The results of analysis by WebProAnalyst display that: (1) for hemagglutinins, amino acids acids, which have most impact on the Pad value are located mostly in HA1 fragment of the protein (positions 1-300), HA2 region has much lower number of such positions; (2) regions with high SADC values or high specificity of amino acid substitutions to high and low Pad cover positions that participate in receptor and sialic acid binding.

Conclusion: The result demonstrates in general applicability of using such characteristics as adhesion index to evaluate various sequence and structural-functional properties of influenza A neuraminidase and hemagglutinins.

Acknowledgement: The work supported by FP7 (FP7-HEALTH-F5-2010-260429), RFBR (11-04-92712), RAS program 6.8, SB RAS project 136.

References:

1. Sachdeva G. et al. (2005) SPAAN: a software program for prediction of adhesins and adhesin-like proteins using neural networks, *Bioinformatics,* **21**: 483-491.
2. Bao Y. et al. (2008) The Influenza Virus Resource at the National Center for Biotechnology Information, *J. Virol.*, **82**: 596-601.
3. Ivanisenko V.A. et al (2005) WebProAnalyst: an interactive tool for analysis of quantitative structure-activity relationships in protein families. *NAR*, **33**:W99-104.

NON-UNIQUENESS OF CYCLES IN GENE NETWORKS MODELS

Akinshin A.A.[1], Golubyatnikov V.P.*[2]
[1] *Altai State Technical University, Barnaul, Russia;*
[2] *Sobolev Institute of Mathematics, SB RAS, Novosibirsk, Russia*
e-mail: glbtn@math.nsc.ru
** Corresponding author*

Key words: *negative feedbacks, periodic trajectories*

Motivation and Aim: We study non-uniqueness of cycles of odd-dimensional nonlinear chemical kinetics dynamical systems which describe regulation of gene networks by negative feedbacks. Numerical and mathematical investigation of periodic regimes of similar gene networks is very important for the computational systems biology.

Methods and Algorithms: Our studies of phase portraits of gene networks models are based on topological methods elaborated in [1], [2], and on complex of computer programs **GeneNetworkModeller** composed by A.A.Akinshin.

Results: We consider symmetric gene networks models represented by nonlinear dynamical systems of the type:

$$(1)\quad \frac{dx_1}{dt}=\frac{\alpha}{1+x_{2k+1}^{\gamma}}-x_1;\qquad \frac{dx_2}{dt}=\frac{\alpha}{1+x_1^{\gamma}}-x_2;\qquad \ldots\qquad \frac{dx_{2k+1}}{dt}=\frac{\alpha}{1+x_{2k}^{\gamma}}-x_{2k+1}.$$

For $k>3$, we detect several cycles of this system, and construct their disjoint neighborhoods. One of them is invariant and contains a stable cycle.

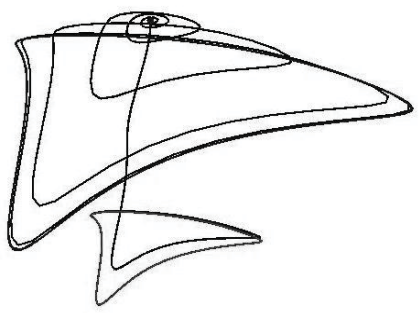


Fig. 1. Two cycles in 9-D symmetric model (1), α=66, γ=5.

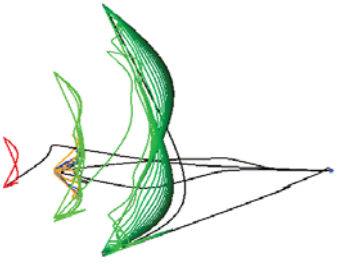


Fig. 2. Three cycles in 15-D symmetric model (1), α=130, γ=6.

Figures show projections of trajectories and limit cycles onto appropriate 2-D planes.

Conclusion: Similar approach can be used in cases of more complicated networks.

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

References:

1. V.A.Likhoshvai, V.P.Golubyatnikov et al. (2008) Theory of gene networks, In: *System computerized biology*, (N.A.Kolchanov, S.S.Goncharov), Novosibirsk, SB RAS, 397-480.
2. V.P.Golubyatnikov, I.V.Golubyatnikov (2011) On periodic trajectories in odd-dimensional gene networks models, *Russ. J. Numer. Anal. Math. Modelling*, 26: 397-412.

UNSTABLE CYCLES IN GENE NETWORKS MODELS

Akinshin A.A.[1], Gaidov Yu.A.[2], Golubyatnikov V.P.*[3], Golubyatnikov I.V. [3]
[1] *Altai State Technical University, Barnaul, Russia;*
[2] *Novosibirsk State Pedagogical University, Novosibirsk, Russia;*
[3] *Sobolev Institute of Mathematics, SB RAS, Novosibirsk, Russia*
e-mail: glbtn@math.nsc.ru
** Corresponding author*

Key words: *negative feedbacks, gene networks models, periodic trajectories, stability*

Motivation and Aim: For odd-dimensional nonlinear dynamical systems modeling gene networks regulated by negative feedbacks, we study non-uniqueness of periodic trajectories and questions of their stability. Such a non-uniqueness and multistablity of gene network models is very important for the computational systems biology.

Methods and Algorithms: Stability, unstability of the cycles, and other properties of gene networks models were established with the help of topological methods elaborated in [1], using computer programs complex **GeneNetworkModeller** composed by A.A.Akinshin.

Results: We consider symmetric gene networks models represented by nonlinear chemical kinetics dynamical systems of the type:

$$(1)\qquad \frac{dx_1}{dt}=f(x_{2k+1})-x_1;\qquad \frac{dx_2}{dt}=f(x_1)-x_2;\qquad \ldots\qquad \frac{dx_{2k+1}}{dt}=f(x_{2k})-x_{2k+1}.$$

Here *f(x)* is a smooth monotonically decreasing function, which represents negative feedbacks in the gene network model. It was shown in [1] that the system (1) has a unique stationary point S_0 surrounded by an invariant domain Q of this system.

We assume that S_0 is a hyperbolic stationary point, and that $2k+1$ is not a prime number: $2k+1=(2m+1)\cdot(2n+1)$. In this case we show that each divisor of $2k+1$ corresponds to an invariant subspace of the system (1), find conditions for existence of a cycle of this system in each of these invariant subspaces, and prove that if $\left|\eta+\frac{df}{dx}\right|<\eta\cdot\sin\frac{2\pi}{2k+1}\cdot\sin\frac{\pi}{2k+1}$ in the invariant domain Q for some positive parameter η, then intersection of each of these subspaces with Q contains a cycle which is stable **in this intersection.** Numerical experiments show that outside of these invariant subspaces the trajectories of the dynamical system (1) converge to its cycle which is stable in the total phase space of this system. Existence of such a stable cycle was established in [1], see also [2].

Conclusion: Similar results can be obtained in the cases of prime values of $2k+1$, and for more complicated models of gene networks, such as in [3].

Acknowledgements: Supported by RFBR grant 12-01-00074, and by SB RAS, Interdisciplinary Projects 80, 136.

References:

1. V.P.Golubyatnikov, I.V.Golubyatnikov (2011) On periodic trajectories in odd-dimensional gene networks models, *Russ. J. Numer. Anal. Math. Modelling*, 26: 397-412.
2. V.P.Golubyatnikov et al. (2010) On the existence and stability of cycles in five-dimensional models of gene networks, Numerical Analysis and Applications. 3: 329-335.
3. 3. Yu.A.Gaidov, V.P.Golubyatnikov (2011) On the existence and stability of cycles in gene networks models with variable feedbacks, Contemporary mathematics, 553**:** 61-74.

ArchiP: DETECTOR OF ARCHITECTURES IN 3D PROTE IN STRUCTURES

Aksianov E.A.*, Alexeevski A.V.
A.N. Belozersky Institute and Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University,
Scientific-Research Institute for System Studies, Russian Academy of Science (NIISI RAN)
e-mail: evaksianov@gmail.com
** Corresponding author*

Key words: β-*sheet, protein architecture, automated detector*

Motivation and Aim: Fold description and classification of 3D protein structures currently requires human judgment in certain cases. The lack of rigorous criteria in fold definitions leads to misunderstanding between different experts in structural classifications. The more difficult step in fold description is characterization of the domain architecture, i.e. spatial arrangement of β-sheets and α-helixes in 3D space. The aim of the work is to elaborate formal definitions of architectures and to develop an automated detector of architecture of all-β or β/α classes according SCOP classification.

Methods and Algorithms: β-sheets are detected by a previously developed program, named SheeP (http://mouse.belozersky.msu.ru/sheep), α-helixes are detected by DSSP algorithm. β-sheets and α-helixes are primary structural units of the architecture. Pair contacts of structural units are determined. Modifications of the set of structural units on the base of contact regions in β-sheets are implemented. Architecture of a protein domain is described in terms of graph, vertices of which are structural units and edges are pairwise contacts between them. The algorithm detects β-sandwiches, β-barrels (open or closed), βα-sandwiches, and six other architectures.

Results: A program ArchiP and web-service were created. To test ArchiP we select a set of domains (one domain from one SCOP family) for each tested architecture. ArchiP detected correctly

– 98.7 % (611 of 619) domains selected from folds, annotated in SCOP as αβα-sandwich folds,

– 93.8 % (245 of 261) domains selected from folds, annotated as β-sandwiches,

– 68.0 % (81 of 119) domains from TIM-barrel folds,

– 62.7 % (140 of 223) domains from β-barrel folds,

– 34.7 % (8 of 23) domains from ββα-sandwich folds; it was observed, that ArchiP adds additional elements into architecture in a number of cases, which actually exist in structure according to expert judgment.

Except ArchiP mistakes, in certain cases misannotations are due to the absence of clear architecture information in SCOP for particular domains and families.

Conclusion: The results of ArchiP demonstrate the possibility of more rigorous approach for protein domain fold description and are of practical usage for protein domain architectures and fold classification.

Availability: Web-service of ArchiP (beta-version) is available at http://mouse.belozersky.msu.ru/archip.

Acknowledgements: The work is partially supported by RFBR grants 10-07-00685-a and 11-04-91340.

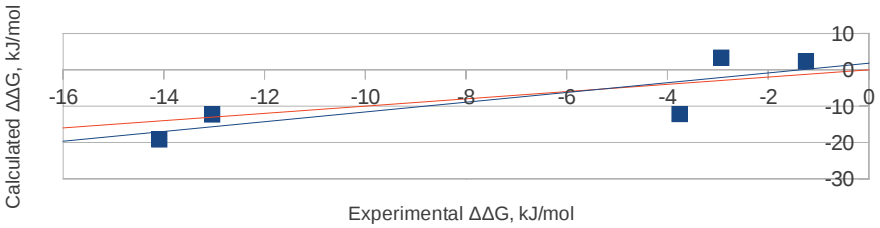
PROTEIN THERMAL STABILITY STUDY USING NAMD ON HIGH-PERFORMANCE CLUSTER

Alemasov N.A.
Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
e-mail: alemasov@bionet.nsc.ru

Key words: *protein thermal stability, NAMD, free energy, high-performance computations*

Motivation and Aim: One of the most exciting challenges in the area of biotechnology and pharmaceutics is design of novel drugs. It has become even more important because of increased viruses activity. The major problem in such investigations is high viruses mutability – often recently developed drugs are no longer effective against viruses they were targeted to. Fortunately high-performance computer simulations can suggest a solution [1].

Methods and Algorithms: There was used widely spread software package called NAMD [2] in the current work to perform molecular dynamics simulation. The main idea of the study is to calculate difference in "alchemical" free energy between wild-type protein and mutated one during their folding.



Results: We suggest the following graph representing comparison between experimental [3] and computational free energy values.

BActerial RiboNucleASE mutants (A32F A32M A32V G34A G34T) were used as a testing example allowing us to verify our approach to calculating protein thermal stability.

Conclusion: We obtained correlation 0.81 than show our approach is viable and able to become a base of industrial methods to distinguishing the most stable virus mutations that developing drugs should target to.

Availability: Scripts and configuration parameters can be available via e-mail to the author.

Acknowledgements: the work is supported by the Ministry of Education and Science state contract #07.514.11.4011.

References:

1. D. Seeliger, B.L. de Groot. (2010) Protein Thermostability Calculations Using Alchemical Free Energy Simulations, Biophys J, 98(10): 2309–2316.
2. J.C. Phillips et al. (2005) Scalable molecular dynamics with NAMD, Journal of Computational Chemistry, 26(16): 1781–1802.
3. M.D. Kumar et al. (2006) ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions, Nucleic Acids Res, 34(Database issue): D204–D206.

SYSTEM BIOLOGY ANALYSIS OF *HELICOBACTER PYLORI* VIRULENCE AND ADAPTATION BASED ON PROTEOGENOMIC, TRANSCIPTOMIC AND METABOLOMIC ANALYSIS

Alexeev D.G., Momynaliev K.T., Selezneva O.V., Demina I.A., Pobeguc O., Tvardovsky A., Altukhov I.A., Govorun V.M.
Research Institute for Physico-Chemical Medicine
e-mail: alexeev@niifhm.ru

Key words: *System biology, proteomics, bacteria, metabolomics*

Motivation and Aim: Helicobacter pylori is a Gram-negative, microaerophilic, helical-shaped bacterium that colonizes the human stomach of at least half of the world's population.*H.pylori* preferentially colonizes the antrum of the stomach, where acid-producing parietal cells are not present, and the environmental pH is higher than in the corpus. In most cases, *H. pylori*can persist in the human stomach asymptomatically, but in some cases, *H. pylori* may progress to symptomatic chronic gastritis, gastric or duodenal ulcers, or gastric cancer. H. pylori is an extra macro- and microdiverse bacterial species. The high level of macrodiversity is supported by the fact that up to 25% of H. pylori genes are dispensable in at least one strain. The most unusual characteristic of the nucleotide sequence diversity of H. pylori is very high number of unique sequences for a given gene across the different strains.

Methods and Algorithms: 3 laboratory strains and one clinical isolate were subject to proteogenomic profiling, with following analysis of transcription and translation by means of PCR and high throughput HPLC-MS/MS analysis. Additionally metabolic potential was measured by metabolite MS profiling.

Results: Clinical isolate was sequenced and compared to laboratory strains. Each strain possesses over 10000 SNP. Over 600 proteins were identified in each strain, however over 100 from each group were uniquely expressed in one of the strains while similar genes silent in the others. Some transcripts were found for the present proteins while for others there were no transcription. System model explaining genetic basis for variability in functional states was built.

Conclusion and Availability: Variable expression is controlled by multiple factors – such as genome SNPs, asRNA, methylation and others. Therefore it allows for high adaptability and diversity. Some insights into functional state associated with high virulence were made.

DEEP METAGENOMICS AND METAPROTEOMICS OF HUMAN GUT: DRAMAS AND DELIGHTS

Alexeev D.G.*, Tyakht A.V., Popenko A.S., Belenikin M.S., Altukhov I.A., Pavlenko A.V., Kostryukova E.S., Selezneva O.V., Larin A.K., Karpova I.Y., Govorun V.M.
Research Institute of Physico-Chemical Medicine FMBA, Moscow, Russia;
Moscow Institute of Physics and Technology, Dolgoprudny, Moscow Region, Russia;
Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow, Russia;
Institute of Bioorganic Chemistry of the RAS, Moscow, Russia;
RRC Kurchatov Institute, Moscow, Russia
e-mail: alexeev@niifhm.ru
** Corresponding author*

Key words: *metagenomics, metaproteomics, human gut microbiota*

Motivation and Aim: Gut microbiota is an essential component of healthy human existence, influencing metabolism, immunity and other facets of organism homeostasis. Composed of multiple microbial species, its metagenomic composition allows to assess the total metabolic potential of such complex ecological system. The study targets examination of 132 metagenomic samples of gut microbiota across a wide range of Russian metropolitan and rural areas using next-generation whole-genome sequencing.

Methods and Algorithms: Phylogenetic and functional profiling of metagenomic samples is performed basing on coverage depth resulting from alignment of reads to catalog of reference sequences, as well as statistical analysis and visualization. The reference sets contain prevalent microbial genomes and genes of human gut microbiota. Samples were compared across various socio-geographic, age- and health-related groups by means of statistical analysis and visualization using R language.

Results: Comparative analysis of Russian samples together with existing large metagenomic data sets from MetaHIT and Human Microbiome Project revealed both significant similarities, as well as novel distinctions across continents on a world scale.

COVERAGE DEPTH ANALYSIS IN NEXT GENERATION SEQUENCING DATA

Amstislavskiy V.S.*, Sultan M., Kim K., Schrinner S., Lehrach H., Yaspo M.-L.
Max Planck Institute for Molecular Genetics, Berlin, Germany
e-mail: amstisla@molgen.mpg.de
** Corresponding author*

Key words: *next generation sequencing (NGS), exome enrichment, whole-genome sequencing (WGS), coverage depth, DNA complexity, copy number variation*

Motivation and Aim: NGS is now a common and widely used approach for a comprehensive analysis of the genetic information in health and disease. Specialized protocols (e.g. WGS, RNA-Seq, Exome sequencing, ChIP-Seq etc.) allow scientists to investigate sequence information at the genome and transcriptome levels. But despite of expanding NGS applications and analysis tools, the biases introduced by the sequencing methods and/or by the context properties (GC-content, complexity) of the genomic regions of interest are not yet fully understood. Herein we developed an interactive web-based tool enabling the online visualization of the distribution of coverage along chromosomes, exons, transcripts or regions of interests for different sequencing applications. This tool allows the visualization of the distribution of sequence reads and helps to estimate the impact of context properties on the data quality through an intuitive web interface. It further helps to identify at a glance copy number variation in e.g. cancer related whole genome sequencing data.

Methods and Algorithms: Raw sequence data underwent quality controls (custom algorithm, fastqc) and mapped using BWA/samtools on the UCSC hg19 reference genome. The coverage profiles are being generated using pileup files and an annotation file (NCBI36, Ensembl 62). The Python matplotlid module was used for graphics and php/mysql - for web-programming.

Results: The tool developed herein was used to analyze the coverage depth of sequence data produced in the context of Oncotrack, a European consortium project aiming at identifying new markers for colon cancer. We used results from WG and mRNA sequencing experiments generated on the Illumina HiSeq2000 platform as well as exome enriched sequence data produced on the SOLiD 5500 platform. We built graphs with simultaneously representation of the average depth of coverage in a given bin as well as the fraction of bin positions covered at least ones. Additionally, the bin GC-content and informational entropy curves were overlaid. We observe that in general the coverage correlates with the informational entropy: it is lower in case of GC-content different than 50%. However, in some regions of WGS sharing the same complexity, we observe significant coverage differences, which are linked to copy number variation events occurring in the tumor sample. We further observe that increasing the number of sequences per experiment by additional sequencing has no drastic effect on the profile of the coverage curve. However enrichment rates in exome protocols and DNA quality significantly affects the average coverage.

Conclusion: The tool will be freely available on the web requiring no prior installation to be used. As an input it requires only a pileup file of the data to be analyzed. It is easy to use, and it is platform independent, which makes it useful for quick checking of the quality of sequencing data, coverage estimation and fast capturing of potential regions of interest for the further analysis.

RECOGNITION OF NFAT5 BINDING SITES

Ananko E.A.*, Levitsky V.G., Efimov V.M., Afonnikov D.A.
Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
e-mail: eananko@bionet.nsc.ru
** Corresponding author*

Key words: *transcription factor binding site, osmotic stress, hypertonic stress*

Motivation and Aim: Transcription factor Tonicity-responsive Enhancer Binding Protein (TonEBP/NFAT5) is involved not only in regulation of kidney homeostasis cell adaptation to osmotic and hypertonic stresses, but also in regulation of myoblast migration and differentiation, cardiac development, and cancer invasion. The aim of current work was to develop the method for NFAT5 binding site recognition and to predict NFAT5 target genes.

Methods and Algorithms: NFAT5 binding sites were compiled from literature. The genetic algorithm was applied for the sequences alignment [1]. Kullback-Leibler Discreate Information Content and P-value were applied to evaluate each DNA motif. Dinucleotide position weight matrix for NFAT5 binding site was constructed as described earlier [2].

Results: The sample of 38 NFAT5 binding sites compiled from literature was used to develop the dinucleotide position frequency and weight matrixes. The length of resulted weight matrix was equal to 22 nucleotides. Figure 1 illustrates conservative positions of the weight matrix.

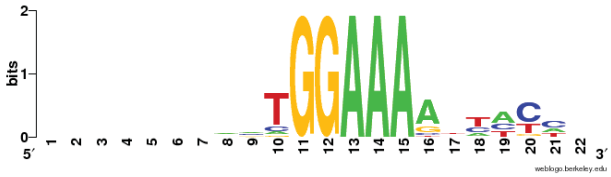


Figure 1. Logo-illustration of the matrix conservative positions.

The method developed was applied for prediction of the NFAT5 target genes in human and mouse genomes. The recognition was made in promoter regions 2000 bp upstream transcription start sites. The results were verified on microarray data published by Lee S.D. and co-authors [3].

Conclusion: The NFAT5 binding sites were predicted in 321 mouse promoters (from 24 531) and in 374 human promoters (from 32 282). Five human genes (*SLC6A12, MAGEA6, TPT1, CCNB3, and ELF1*) are of particular interest. Promoter regions of these genes contain two different putative NFAT5 binding sites.

Acknowledgements

The work was supported in part by SB RAS (Integration project No. 136), RAS (projects Nos. 6.8, 30.29), Russia's President's project No. 5278.2012.4 (scientific school), Russian Ministry of Education and Science (projects Nos. 07.514.11.4023, 02.740.11.0882).

References:

1. V.G. Levitsky. (2012) DNA Motif search by genetic algorithm, Proceedings BGRS 2012
2. V.G. Levitsky, E.V. Ignatieva, E.A. Ananko 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.
3. S.D. Lee, S.Y. Choi, S.W. Lim et al. (2011) TonEBP stimulates multiple cellular pathways for adaptation to hypertonic stress: organic osmolyte-dependent and -independent pathways. Am. J. Physiol. Renal. Physiol., 300: F707-F715.

THE REVIEW OF EXISTING SERVICES IN THE FIELD OF PERSONALIZED GENETICS

Anashkin S.S.
Company executive, "Genoanalytica" JSC, Moscow, Russia
e-mail: sa@genoanalytica.ru

The demand for study of personal DNA in its infancy. Most of the companies involved in the promotion of DTCGT, rely on the curiosity of consumers and social networks on the Internet. Physicians are traditionally more conservative in everything new, they know not very much about genetics. They have no experience in the application of the results for research DNA in their clinical practice.

Market DTCGT is being born for now. At this stage it is important to assess the willingness of society to consume directly, bypassing a doctor. On the other hand, the question is whether this expediency?

With the use of PCR methods. It's almost all the major network laboratories Center for Molecular Genetics. The products of these laboratories in different combinations and with different names are in the range of services many commercial hospitals. Available analyzes of polymorphisms of 1 to several, batch studies on core areas: cardiovascular, oncological, endocrinology, neurodegenerative, autosomal. Prices range from 150 rubles for the polymorphism of up to tens of thousands of batch studies. Some medical centers integrate such research into their own programs "check-up" diagnostics.

Microarrays. There was about 3 years ago, which allowed to talk about that a few years ago, VS Baranov called the "genetic passport". Screening study on several thousands of SNP, which is the main feature of the method of chip, as a rule, artificially divided into the following packages:

-Monogenic disease (carrier status)
-multifactorial (polygenic) diseases (predisposition)
-Pharmacogenetics (reaction to medications)
-physical characteristics/specifications (+ sports genetics)

The cost of such "wholesale" method of research (thousands of SNP at once) as a result is cheaper PCR method.

However, managers of laboratories that perform conventional PCR assays, insist that the patient does not need it all at once, the entire screening. Consequently, consumers do not need to pay from 30 000 to 75 000rub - a range of cost studies of DNA on the microarray. From this we can agree, provided the consumer-patient, but rather, his doctor knows exactly - what specific SNP should be analyzed. In all other cases it is advisable to conduct a one-time screening of the maximum number of SNP.

We should also provide a completely new high-tech area - sequencing of individual exons or entire genes, or even a few genes at once. And although so far the cost of such studies is still high, but there are technological solutions that able significantly affect the price. In this regard, very soon, probably, it will be possible to note the appearance of the new market of DNA testings, which will make adjustments to the current structure of the market.

What motivates consumers of DTCGT:

- Cognitive, curiosity (fun)
- Responsible for managing health
- Planning for pregnancy, IVF (carrier status of monogenic diseases)
- Prescription of doctor to confirm or refute the diagnosis, or for adjustments of treatment
- Common sense (the possession of the knowledge available today to help with good reason and consciously choose a lifestyle that is harmonious own genes)

About genetic discrimination (in Russia there is no law); on ethical issues.

THEORETICAL STUDY OF STRUCTURAL FEATURES OF VARIOLA VIRUS CrmB PROTEIN

Antonets D.V.*, Nepomnyashchikh T.S., Shchelkunov S.N.
State Research Center of Virology and Biotechnology "Vector", Koltsovo, Russia
e-mail: antonec@yandex.ru
** Corresponding author*

Key words: *tumor necrosis factor, orthopoxvirus, viral immunomodulatory proteins, CrmB, molecular modeling*

Motivation and Aim: Orthopoxviral TNF-binding proteins and especially variola virus (VARV) CrmB may be used to develop novel medications for treatment of rheumatoid arthritis, Chron's desease and other pathologies driven by TNF overproduction. The aim of this study was the theoretical analysis of molecular mechanisms underlying interaction of orthopoxviral TNF-binding CrmB proteins with their ligans.

Methods and Algorithms: Models of TNF receptor domains of VARV- and CPXV-CrmB and their complexes with different TNFs were constructed using Modeller (9v2) software (http://salilab.org/modeller). All constructed models were then energy minimized using either NOC (http://noch.sourceforge.net) or FoldX (http://foldx.crg.es). Stability of ligand-receptor complexes was predicted either with FoldX or using residue-level pairwise potentials BETM990101. FoldX was used for designing mutant forms of VARV CrmB. Spatial structure of VARV-CrmB C-terminal domain was predicted with I-TASSER (http://zhanglab.ccmb.med.umich.edu/I-TASSER/).

Results: Analysis of produced ligand-receptor models with either FoldX or with BETM990101 pair potentials revealed that mTNF should bind to CPXV-CrmB with higher affinity than hTNF. VARV-CrmB was predicted to bind both cytokines with higher affinity than CPXV-CrmB; CPXV-CrmB was predicted to bind hTNF(R31Q) with significantly higher affinity than wild type hTNF. And both CrmBs were predicted to less efficiently bind to hTNF(E127Q), than to the wild type hTNF. All these findings were then qualitatively approved by experimental evaluation of VARV- and CPXV-CrmB proteins ability to inhibit cytotoxic action of mTNF, hTNF, hTNF(R31Q) and hTNF(E127Q) on L929 murine fibroblast cells. These models were then used for designing mutant forms of VARV-CrmB which should have higher affinity to hTNF. Several mutant forms of CrmB, predicted with FoldX to be the most affine to hTNF, were chosen for further theoretical analysis and experimental evaluation.

Using the I-TASSER the spatial structure of VARV CrmB C-terminal chemokine-binding domain (SECRET) was predicted and it was assumed to be the structural homologue of CPXV vCCI protein belonging to the family of poxviral type II chemokine-binding proteins despite weak homology of their amino acid sequences (12 %). We suggested that SECRET should be included into the family of poxviral type II chemokine-binding proteins and that it might have been evolved from the vCCI-like predecessor protein. Recently our predictions were confirmed by the X-ray structure of Ectromelia virus CrmD protein SECRET domain.

Acknowledgements: This work was supported by Russian Foundation for Basic Research grants #10-04-00479-a and #12-04-00110-a.

TEpredict – SOFTWARE FOR PREDICTING T-CELL EPITOPES. AN UPDATE

Antonets D.V. *, Grudin D.S.
State Research Center of Virology and Biotechnology "Vector", Koltsovo, Russia
e-mail: Den.Antonets@gmail.com
** Corresponding author*

Key words: *T-cell epitope, immunoinformatics, machine learning*

Motivation and Aim: CD8+ T-cell epitopes play crucial role in antiviral and anticancer immunity and accurate *in silico* identification of potent T-cell epitopes could drastically reduce materials and time consumption compared to the traditional experimental approaches of epitope discovery. The main aim of this work was the development of new models for predicting peptide binding to different allomorphs of MHC class I molecules to update our TEpredict software [1].

Methods and Algorithms: New models for predicting affinity of peptide-MHC binding were constructed by means of either partial least squares (PLS) or with recently developed sparse partial least squares (SPLS) technique or using random forest algorithm (RF). Peptide:MHC binding data for producing the models was collected from Immune Epitope Database (IEDB; http://www.immuneepitope.org). Recently developed amino acid similarity matrix PMBEC, derived from experimentally determined peptide:MHC binding data [2], was used here to parameterize amino acid residues along with sparse encoding. Using independent component analysis (ICA) [3] of PMBEC matrix several amino acid parameterization schemes with reduced dimensionality were produced. Performance of generated models was assessed with ROCR package [4]. All programs were written in Python programming language; Python interacted with R through the RPy2 interface (http://rpy.sourceforge.net/).

Results: Models built using PMBEC parameterization of amino acid residues were shown to outperform those built with sparse-encoding. All PLS, SPLS and RF-based models built using ICA-based scales with dimensionality of 11 were shown to slightly outperform sparse-encoding-based models and PMBEC-based ones both in terms of the area under the ROC curve and Pearson's correlation coefficient.

Availability: TEpredict could be freely downloaded at http://tepredict.sourceforge.net.

Acknowledgements: This work was supported by the Federal Target Program "Research and development on priority directions of scientific-technological complex of Russia for 2007-2012" (contract#16.512.11.2186).

References:

1. D.V. Antonets, A.Z. Maksyutov (2010) TEpredict: software for T-cell epitope prediction. *Mol. Biol.*, **44**:119-127.
2. Kim Y. et al. (2009) Derivation of an amino acid similarity matrix for peptide: MHC binding and its application as a Bayesian prior. *BMC Bioinformatics*, **10**:394.
3. Karvanen J., Koivunen V. (2002) Blind separation methods based on Pearson system and its extensions. *Signal Processing,* 82:663–673.
4. Tobias S. et al. (2005) ROCR: visualizing classifier performance in R. *Bioinformatics*, **21**: 3940-3941.

COMPARING HOEFFDING'S D MEASURE AND MAXIMAL INFORMATION COEFFICIENT FOR ASSOCIATION ANALYSIS

Antonets D.V.*[1], Cheryomushkin E.S.[2], Vyatkin Yu.V. [3]
[1] *State Research Center of Virology and Biotechnology "Vector", Koltsovo, Russia;*
[2] *Ershov Institute of Informatics Systems, Novosibirsk, Russia;*
[3] *AcademGene LLC, Novosibirsk, Russia*
e-mail: antonec@yandex.ru
** Corresponding author*

Key words: *association analysis, Hoeffding's D, maximal information coefficient*

Motivation and Aim: Discovering dependencies within the data is very important for analyzing gene expression, reconstructing genetic networks and for numerous other data-rich applications. Recently Reshef et al. [1] described novel maximal information coefficient (MIC) aimed to measure dependence between two variables. The main advantages of MIC, as reported, are its generality – the ability to capture many kinds of relationships, not limited to specific function types – and equitability – ability to give similar scores to equally noisy relationships of different types. In the paper this criterion was also used to analyze gene expression data. Although MIC was shown by the authors to outperform either Pearson's or Spearman's correlation (currently the most common choices for analyzing gene expression data) and several other methods, we were interested to compare MIC with another powerful technique, namely Hoeffding's D measure developed in 1948 [2], that was also recently confirmed to be efficient in this field [3].

Methods and Algorithms: For calculating MIC we used MINE utility available at http://www.exploredata.net. Corresponding p-values for MIC were either taken from the site or calculated by us. D measure was calculated using function *hoeffd* in Hmisc package for R. We analyzed different kinds of associations varying the number of samples (30, 100 and 1000), the level of noise (0, 10, 20 and 30 %) and outliers ratio (5, 10 and 15 %). All simulations were done using R (www.r-project.org). We also applied both criteria for analyzing gene expression in embryonic development of *Drosophila melanogaster*.

Results and conclusion: Both criteria outperformed either of correlation methods, but MIC was found to be more sensitive to noise and outliers and to require more samples to draw strong conclusions on dependency as compared to Hoeffding's D, and its "equitability" was strongly dependent on sample size. Hoeffding's D was found to be equally efficient as MIC in analyzing gene expression data. Besides MIC, MINE utility produces some other useful metrics of the data [1], but given that biological data is often sparse, noised and contains outliers, MINE, if applicated, should be accompanied by Hoeffding's D measure.

References:

1. Reshef D.N. et al. (2011) Detecting novel associations in large data sets, *Science*, **334**:1518-1524.
2. Hoeffding W. (1948) A non-parametric test of independence, *Annals of Mathematical Statistics*, **19**:293–325.
3. Fujita A. et al. (2009) Comparing Pearson, Spearman and Hoeffding's D measure for gene expression association analysis, *J Bioinform Comput Biol*, **7**:663-684.

MUTATIONS IN *K-RAS* AND *EGFR* GENES AND THE SEARCH FOR SNPs, ASSOCIATED WITH THEIR OCCURRENCE

Antontseva E.V.*[1], Bryzgalov L.O.[1], Matveeva M.Yu.[1], Ponomaryova A.A.[2], Ivanova A.A.[2], Rykova E.Y.[3], Cherdyntseva N.V.[2], Merkulova T.I.[1]
[1] *Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia;*
[2] *Cancer Research Institute SB RAMS, Tomsk, Russia;*
[3] *Institute of Chemical Biology and Fundamental Medicine SB RAS, Novosibirsk, Russia*
e-mail: schly@mail.ru
** Corresponding author*

Key words: *K-Ras, EGFR, somatic mutations, single nucleotide polymorphism (SNP) , lung cancer*

Motivation and Aim: The purposes of this research were to obtain data on the frequency of somatic mutations in *EGFR* and K-*Ras* genes and inherited SNPs in *K-Ras* among patients with non-small cell lung cancer (NSCLC) in the Western Siberian region of Russia, the development of minimal invasive diagnostics for these mutations by the use of extracellular DNA, analysis of potential rSNP in *K-Ras* gene, and the search for new rSNP, which may be relevant to the development of NSCLC.

Methods and Algorithms: DNA for genotyping was isolated from 40 NSCLC tumors, 14 of them identified as adenocarcinoma (AC), and 26 as squamous cell carcinoma (SCC) Detection of the somatic mutation Leu858Arg in *EGFR*, was carried out by allele-specific RT-PCR with SybrGreen. Identification of 9-18 bp microdeletion of *EGFR* gene (the region of 746-750 amino acids of the protein) was performed by PCR with further separation by gel electrophoresis in agarose gel. Mutations at the codon 12 in *K-Ras* gene was determined by sequencing after enrichment of DNA template with mutant allele. A search for potential rSNPs in *K-Ras* intron 2 was carried out by sequencing of patient's DNA to identify markers of genetic susceptibility to lung cancer. The electrophoretic mobility shift assay was used to study the predicted rSNPs.

Results: Mutations in the *EGFR* gene were found in 10 % of ACs and in 7 % of SCCs. Mutations at codon 12 in *K-Ras* were identified in 29 % of ACs and in 8 % of SCCs. Tumors with simultaneous presence of mutations in both genes were not detected in this sample. We identified two SNPs in intron 2 in *K-Ras*, affecting the binding sites of transcription factors (TF) NF-Y and GATA-6 and characterized by an increased frequency of occurrence (3.3-3.5-fold) relative to the random sample.

Conclusion: Despite the low volume of the studied sample, we can make a preliminary conclusion that in patients with NSCLC of the West Siberian region, the frequency of mutations studied do not differ from the average, and cases of joint detection of two mutations are very rare. We have identified two rSNPs in intron 2 of *K-Ras* that affect binding sites for the same TF as in the case of mice [1] in which similar SNPs have a *cis*- effect on the formation of mutations at codon 12 K-ras.

References:

1. E.V. Gorshkova et al. (2006) Lung cancer-associated SNP at the beginning of mouse K-ras gene intron 2 is essential for transcription factor binding, *Bull Exp Biol Med,* **6:** 731-733.

PATHWAY SIGNAL FLOW ANALYSIS FOR HIGH-THROUGHPUT GENE EXPRESSION DATA

Arakelyan A.A.
Institute of Molecular Biology NAS RA, Yerevan, Armenia
e-mail: aarakelyan@sci.am

Key words: *high-throughput gene expression, data analysis, molecular pathway mining*

Motivation and Aim: Molecular pathway sources (such as KEGG and Biocarta) may become the most useful tools for gene expression analysis. To date, in most cases pathway mining simply means mapping and coloring nodes in pathways without evaluation of changes of flows through the network. Herein, we introduce a new algorithm for functional annotation and biological interpretation of gene expression data based on in-depth pathway analysis.

Methods and Algorithms: The Pathway Signal Flow (PSF) algorithm evaluates how a signal from network inputs spreads downstream to the outputs depending on relative expressions of nodes (R) and interaction types between them. The signal flow between two connected nodes is calculated as:

$R_{(node1)} * R_{(node2)}$, *if node 1 activates node 2, and* $R_{(node\ 1)}/R_{(node\ 2)}$, *if node 1 inhibits node 2*.

The PFS value is the mean of signal flows at output nodes. Empirical probability distribution of PSFs is used to calculate significance of PSF values.

Results: We have reanalyzed data from several microarray datasets related to pulmonary sarcoidosis and compared with published results. Analysis of gene expression was performed using growing support sets algorithm [1] combined with PSF analysis. Results indicate that sarcoidosis is characterized by up-regulation of pathways related to pro-inflammatory response, such as Fc (IgG) receptor mediated phagocytosis, focal adhesion, chemokine signaling and T-cell receptor signaling. Eminently, among the different possible functional consequences of pathway activation, PSF was able to detect biologically meaningful outcomes. For example, chemokine signaling pathway may induce different responses in cells such as activation and migration of immune cells, expression of pro- or anti-inflammatory mediators, apoptosis. In sarcoidosis, we detected up-regulation of chemokine pathway branches related to cell proliferation and migration, which is in agreement with previously reported experimental results [2].

Conclusion: The PSF analysis is an attempt to achieve deeper level of biological interpretation of gene expression data and provide more biologically meaningful results compared to existing techniques. Moreover, our algorithm is suited for all current techniques for high-throughput gene expression experiments.

Availability: MATLAB code for PSF analysis is available upon request from authors.

References

1. V. Palchevskiy et al. (2011) Immune response CC chemokines CCL2 and CCL5 are associated with pulmonary sarcoidosis, *Fibrogenesis Tissue Repair,* **4:** 10.
2. A. Arakelyan et al. (2010) Growing support set systems in analysis of high-throughput gene expression data, In: *New Trends in Classification and Data Mining*, K. Markov et al (eds)**,** 47-53, ITHEA.

EXPERIMENT-BASED VALIDATION OF COMPUTATIONAL MODELS OF PYRIN – FAMILIAL MEDITERRANEAN FEVER PROTEIN

Arakelyan A.A.*, Nersisyan L., Avetisyan N., Martirosyan G.
Institute of Molecular Biology NAS RA, Yerevan, Armenia
e-mail: aarakelyan@sci.am
** Corresponding author*

Key words: *protein structure modeling, protein-protein docking, structure validation*

Motivation and Aim: One of current issues confronted by computational modeling science is the validation of predicted protein models. Molecular mechanics and knowledge based potentials are not absolute indicators of protein model quality. This paper describes validation of previously obtained structural models of pyrin, a protein implicated in the development of Familial Mediterranean fever, based on their correspondence to available experimental data on pyrin multimerization, interaction with other proteins and phosphorylation.

Methods and Algorithms: Homotrimers of pyrin, as well as its complexes with ASC and caspase-1 proteins were constructed with ClusPro server. Reliability of docked complexes was validated based on empirical cumulative distribution of cluster sizes using kernel smoothing density estimate. Structure based prediction of phosphorylation sites was performed with Phos3D web server.

Results: Docking results indicate that pyrin forms a homotrimer with high significance through its coiled-coil domain. Interactions of pyrin with ASC through its PYD domain and with caspase-1 through its B30.2 domain appear among top 5 complexes obtained by docking. All these results are in good accordance with available experimental data [1]. Phosphorylation site prediction results indicate that residues 208 and 242 are prone to phosphokinase A mediated phosphorylation, whereas residue 209 is not susceptible to phosphorylation, thus showing good agreement with the study conducted Jeru et al [].

Conclusion: Experiment-based validation of 3D structural model of pyrin points on its native-like structure and validity for using in further investigations about structural and functional features of pyrin in its native and mutated states.

Acknowledgements: The work is supported by grant from SCS MES RA (11B-1F014).

References:

1. JW Yu et al. (2007) Pyrin activates the ASC pyroptosome in response to engagement by autoinflammatory PSTPIP1 mutants, *Molecular cell,* **28(2)**: 214-227.
2. N Richards et al. (2001) Interaction between pyrin and the apoptotic speck protein (ASC) modulates ASC induced apoptosis, *The Journal of biological chemistry*, **276(42):** 39320-39329.
3. JJ Chae et al. (2006) The B30.2 domain of pyrin, the familial Mediterranean fever protein, interacts directly with caspase-1 to modulate IL-1beta production, *PNAS USA*, **103(26):** 9982-9987.
4. I Jeru et al. Interaction of pyrin with 14.3.3 in an isoform-specific and phosphorylation-dependent manner regulates its translocation to the nucleus, *Arthritis and rheumatism,* **52(6)**: 1848-1857.

EVALUATION OF GENOMIC INSTABILITY IN SEVERAL SPECIES OF MAMMALS USING THE MICRONUCLEI TEST

Astafieva E.E.*, Karpushkina T.V., Kulikova K.A., Glazko T.T.
Russian State Agrarian University – Moscow Agricultural Academy named after K.A. Timiryazev (RSAU–MTAA), Moscow, Russia
e-mail: k-astafeva@mail.ru
** Corresponding author*

Key words: *micronuclei, genomic instability, species, environmental factors*

Motivation and Aim: Micronuclei test in somatic cells allows to evaluate the genomic instability and predict the reproductive "success" of humans and animals [1, 2]. It is widely used for bioindication of genotoxic environmental effects. However, the species specific traits of the micronuclei test results, the influence of duration of environmental factor action still remain insufficiently studied. In this regard the comparative analysis of the micronuclei test in erythrocytes of the peripheral blood of domesticated animals (cattle, sheep, goats, horses), semi domesticated (yaks) and wild species (musk ox) was carried out. Groups of cattle, sheep, goats, horses, musk ox reproduced in relatively favorable environmental conditions and in the area of high-risk animal breeding of the Southern Gobi Desert (Mongolia) had been compared.

Methods and Algorithms: Blood smears were prepared by mixing a drop of peripheral blood with a drop of saline solution (1:1) on a glass slide then spreading it over a glass slide. Preparations were fixed with methyl alcohol and then stained with Giemsa (Merk). The number of erythrocytes with micronuclei were calculated under a microscope Motik (DMBA300) with a built-in digital camera (x1000) in 3000 cells and expressed in ppm (‰). Statistical significance was evaluated by Student's test (t_S).

Results: The frequency of erythrocytes with micronuclei was significantly higher ($P < 0.05$) in domesticated species compared to the wild species (musk ox). The increase is observed in the following order: musk ox (0.3 ± 0.2 ‰), horses (2.4 ± 0.1 ‰), yak (3.2 ± 0.6 ‰), cattle (4.6 ± 0.7 ‰), goats (4.6 ± 0.4 ‰), sheep (5.3 ± 0.4 ‰). The lowest values of the frequency of erythrocytes with micronuclei were detected in musk oxen, the highest – in sheep (birliksky type of the edilbai sheep – 5.2 ± 0.2 ‰, the Kalmyk breed – 4.3 ± 0.3 ‰, Mongolian sheep – 5.3 ± 0.4 ‰). There were no significant differences in the interbreed micronuclei test results, but animals of different species reproduced in the conditions of zone of high-risk animal breeding (South Gobi, Mongolia) had the statistically significant lower levels of the frequency of red blood cells with cytogenetic anomalies (yaks – 0.3 ± 0.2 ‰, cattle – 1.8 ± 0.6 ‰, goats – 0.9 ± 0.2 ‰, sheep – 0.9 ± 0.1 ‰) compared to the other animals reproduced in relatively more favorable conditions.

Conclusion: A trend to relatively increased genomic instability in domesticated animals compared to the wild ones was revealed. Micronuclei test results indicated also that in the zone of high-risk animal breeding in Southern Gobi Desert the animals with relatively higher genomic stability achieve benefits for the reproduction in generations.

References

1. L. Migliore et al. (2006) Relationship between genotoxicity biomarkers in somatic and germ cells: findings from a biomonitoring study, *Mutagenesis*, Vol. 21. No. **2**: 149-152.
2. J. Rubes et al. (1991) Somatic chromosome mutations and morphological abnormalities in sperms of boars, *Hereditas*, **115**: 139-143.

CIRCULATING microRNAs AS POTENTIAL BIOMARKERS OF LUNG CANCER

Aushev V.N.*[1, 2], Akselrod M.E.[1], Zborovskaya I.B.[1], Krutovskikh V.A.[2]
[1] *Laboratory of Oncogenes Regulation – Institute of Carcinogenesis of N.N. Blokhin Russian Oncological Science Center RAMS – Moscow, Russia;*
[2] *Epigenetics group – The International Agency for Research on Cancer – Lyon, France*
e-mail: vaushev@gmail.com
** Corresponding author*

Key words: *microRNAs, biomarkers, cancer*

Motivation and Aim: The work was devoted to the research of microRNAs as potential cancer biomarkers in blood. We supposed to investigate the limited and possibly homogenous panel of samples from patients with squamous cell carcinoma. In order to further decrease individual variations, design of our study was based on the comparison of miRNA expression profiles of the same patient before and after removal of tumour.

Methods and Algorithms: Patients diagnosed with squamous cell carcinoma underwent surgical resection of the tumour, and blood samples were taken at 2 points: 1 day before surgery, and 7-10 days after. Blood samples were taken in EDTA-containing tubes and centrifuged to separate plasma fraction. RNA was isolated using "NucleoSpin miRNA Plasma" columns from Macherey-Nagel (Germany).

First step included global analysis of a wide set of microRNAs. This initial list of candidate miRNAs (90 miRNA species) was composed based on known data from literature: we included all miRNAs that were reported as possible circulating biomarkers of non-small cell lung cancer, most of miRNAs assigned for other types of lung cancer, and part of miRNAs reported for non-lung cancers. MiRNA profiling was performed by qPCR analysis on custom 384-well microRNA PCR panels from Exiqon (Denmark). Most promising candidates were further validated by qPCR with individual miRNA probes on wider set of samples.

Results: Most of miRNAs from the designed list were successfully detected (with Ct values between 18 and 35) in almost all samples. Highest signal was detected for miR-451, -223, -15a, -486-5p, -16, -21, which is in agreement with literature data. Small part of miRNAs (including miR-206, -518b, -422a, -202, -566) could not be detected and probably are not presented in plasma. We selected miRNAs which displayed the most significant difference between pre- and post-operative samples. Namely, expression of miR-205, -19a, -19b, -451 and -30b was decreased after removal of tumour. Part of these miRNAs were further validated using wider panel of samples. Results of the validation confirm indicated changes of miRNA expression. Additional experiments revealed that most of detected miRNAs are presented in exosomes-enriched fraction of plasma.

Conclusion: Our results suggest that tumour-related miRNAs can be successfully detected in plasma of lung cancer patients and can thus serve as potential biomarkers for this disease.

Acknowledgements: Authors are grateful to colleagues in Oncological Science Center for their assistance in blood samples collection.

THE COMET-FISH TECHNIQUE FOR MONITORING CANCER TREATMENT RESPONSE AT THE GENOMIC LEVEL

Babayan N.S.*, Gevorkyan A.L., Aroutiounian R.M., Hovhannisyan G.G.
Yerevan State University, Yerevan, Armenia
e-mail: babayannelly@yahoo.com
** Corresponding author*

Key words: *personalized medicine, Comet-FISH, telomere, peptide nucleic acid, cancer, cisplatin, bleomycin*

Motivation and Aim. Nowadays, new molecular methods and techniques are developing which can contribute to the objective of personalized medicine [1]. Although telomerase and telomeres have become an attractive therapeutic cancer targets, since most human cancer cells (85-90%) typically express high levels of telomerase [2]. The aim of the present study is the development of a method for comparative investigation of action of widely applied anticancer preparations: cisplatin (cis-DDP) and bleomycin (BLM) on total DNA and telomeres in human blood cells.

Methods. The "Comet-FISH technique" - single cell gel electrophoresis ("comet assay") in combination with fluorescent *in situ* hybridization (FISH) was used for this purpose. This newly applied combined approach permits to detect on the same specimen the total DNA damage in individual cells and evaluate specific DNA sequences as well. Telomere - specific - PNA (peptide nucleic acid) probes were used for the localization of telomeres in the comet's head and their migration to the tail.

Results. By comparing the slopes of the linear regressions of the BLM and the BLM-cis-DDP treatment a slope of 1.07 was found for the BLM treatment alone, which indicates that the telomeres are of the same sensitivity as the average DNA. In contrast the treatment with the combination of BLM and cis-DDP reduces telomere migration more than the migration of total DNA and results in a slope smaller than 1 ($b = 0.77$) and hints to an enhanced cross-linking effect on the telomere sequences. Thus, preferentially telomeric action of the cis-DDP can be concluded.

Conclusion. The presented Comet-FISH approach with telomere PNA permits direct and precise detection of the telomere migration from the former cell nucleus to the comet tail in cells treated with cytostatics, with a direct analysis of correlation to the overall DNA fragmentation. That can be important for monitoring the application of clinical relevant cytostatics during therapy, especially in combinatory approaches, where more than one substance is used at a time.

Acknowledgements. This study was greatly enhanced by the cooperation with prof. E. Gebhart (Institute of Human Genetics, Erlangen, Germany) and prof. A. Rapp (Institute for Molecular Biotechnology, Jena, Germany).

References

1. C.T. Caskey. (2010) Using genetic diagnosis to determine individual therapeutic utility, Annu. Rev. Med., 61: 1–15.
2. A. Rцth *et al*. (2010) Imetelstat (GRN163L) - telomerase-based cancer therapy, In: Small Molecules in Oncology (Recent Results in Cancer Research), U.M. Martens (Edr.), 221-234 (© Springer-Verlag).

HEMAEXPLORER WEBSERVER: VISUALIZATION OF GENE EXPRESSION IN THE HEMATOPOETIC SYSTEM

Bagger F.O.*[1-3], Rapin N.[1–3], Theilgaard-Mönch K.[1, 2, 4], Kaczkowski B.[1, 3], Jendholm J.[2, 3], Winther O.[5], Porse B.[1-3,6]

[1]Bioinformatics Centre, Department of Biology, University of Copenhagen, Copenhagen, Denmark; [2] The Finsen Laboratory, Rigshospitalet, Faculty of Health Sciences, University of Copenhagen, Copenhagen, Denmark; [3] Biotech Research and Innovation Center (BRIC), University of Copenhagen, Copenhagen, Denmark; [4] Dept. of Hematology, Skanes University Hospital, University of Lund, Sweden; [5] Informatics and Mathematical Modelling (IMM), Technical University of Denmark (DTU), Kgs. Lyngby, Denmark; [6] Danish Stem Cell Centre (DanStem), Faculty of Health Sciences, University of Copenhagen, Denmark
e-mail: frederik@binf.ku.dk
** Corresponding author*

Key words: *myeloid leukemias and dysplasias; gene expression profile webtool; gene expression; microarray; webserver*

Motivation and Aim: Researchers within the field of hematopoetic system and Acute Myeloid Leukaemia (AML) have been unable to make use of public available data on an everyday basis, e.g. to look up genes encountered in research or literature. We present a webserver comprising a manual curated database of AML cells as well as normal human hematopoietic stem cells and hematopoietic progenitor cells with an easy interface to get an overview of the mRNA expression profiles.

Methods and Algorithms: The database was build using manually curated public available gene expression data sets from multiple studies all generated on the Affymetrix platform. In order to correct for batch effect the data sets were sorted by platform and laboratory, prior to RMA normalization and subsequent processing with ComBat [1]. Fold changes for AML samples were computed using a novel normalization method that first identifies the nearest normal counterpart to the individual AML sample and use this to compute gene expression changes in the AML sample (N.R., J.J., K.T.M., O.W. and B.T.P, unpublished data). For genes detected by more than one probe sets on the microarrays, gene expression levels are presented by the probe set with the highest mean expression level; data from all available array probe sets are available for download.

Results: The HemaExplore [2] webserver can take one gene as query and provides a plot of the expression of the gene in both hematopoietic stem and progenitor populations as well as in mature lineages. Alternatively, a query of two genes depicts their relationship in a scatter plot. Currently the database contains options for the human hematopoietic system, human AML, and the murine hematopoietic system.

Conclusion: The HemaExplorer webserver will provide researchers with a powerful tool to check expression levels for genes of interest in normal hematopoiesis as well as AML.

Availability: The webserver is freely available at: http://servers.binf.ku.dk/hemaexplorer/

References

1. W.E. Johnson, et al. (2007) Adjusting batch effects in microarray expression data using empirical bayes methods. Biostatistics. 8:118-27.
2. 2. Frederik Otzen Bagger and Nicolas Rapin et al. HemaExplorer: A webserver for easy and fast visualization of gene expression in normal and malignant hematopoiesis. Blood. Accepted for publication.

EFFECT OF CHRONIC O-AMINOAZOTOLUENE TREATMENT ON XENOSENSORS CAR, PPARa, PPARg GENES AND THEIR TARGET GENES EXPRESSION IN MICE CC57BR/Mv AND DD/He

Baginskaya N.V.*, Kashina E.V., Shamanina M.Yu.
Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
e-mail: bagin@bionet.nsc.ru
** Corresponding author*

Key words: *Quantitative real-time RCR, nuclear receptors, hepatocarcinogenesis, mice*

Motivation and Aim: Chemical carcinogens interact with xenosensor nuclear receptors (NR) and effect on their expression and activity. It has been established that changes in NR expression pattern and activity themselves can serve as carcinogenic factors. The aim of our study was to reveal the interaction between sensitivity to chemical carcinogenesis and changes in liver NR expression during chronic hepatocarcinogen application.

Methods and Algorithms: Male mice of DD/He strain (DD) which are sensitive to chemical hepatocarcinogenesis and male mice of CC57BR/Mv (BR) resistant strain were the object of our investigation. O-aminoazotoluene (OAT) was injected intraperitoneally at concentration 225 mg/kg for 4 times during two month. Control animals received injections of solvent (olive oil). Animals were killed in 1 and 4 days after the last injection. *Car, Ppara, Pparg, Ugt1a1* and *Sult1a1* genes expression in the liver were measured by quantitative real-time PCR. mRNA for beta-actin was used as internal control. The inflammation level was estimated by TNF-alpha serum concentration.

Results: The levels of *Car*, *Pparg* and *Ugt1a1*genes expression were significantly higher in the liver of control BR mice as compared with control DD mice, although *Sult1a1* expression level was lower. One day after 4-th OAT injection a considerable increase in serum TNF-alpha level was registered in both strains. Similarly, hepatic mRNA levels of all genes examined in BR strain decreased dramatically during the inflammation while only *Sult1a1* decrease was observed in DD mice. At 4 days after OAT injection, TNF-alpha level in BR mice returned to normal, whereas in DD mice it remained increased. The expression level of *Ppara* was decreased and *Pparg* was increased in this time in both stains compared to control. It may reflect the existence of processes directed to inflammation suppress.

Conclusion: Thus, it has been established that chronic OAT treatment results in short-time inflammation, which followed by decreased in *Car*, *Ppars, Sult1a1* and *Ugt1a1* genes expression in resistant BR mice. In contrast, sensitive DD mice demonstrate longer inflammation process followed by the same or slightly increased *Car*, *Ppars* и *Ugt1a1* gene expression.

Research is supported by RFBR grant №11-04-00545.

MOLECULAR MODELING OF CYTOSOLIC PART OF a2-SUBUNIT OF MOUSE V-ATPase

Bakulina A.*[1], Merkulova M.[2], Hosokawa H.[2], Phat Vinh Dip[3], Gruüber G.[3], Marshansky V.[2]

[1] *State Research Center of Virology and Biotechnology "Vector", Koltsovo, Russia;*
[2] *Center for Systems Biology, Program in Membrane Biology, Simches Research Center, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA;*
[3] *Nanyang Technological University, Division of Structural and Computational Biology, School of Biological Sciences, Singapore*
e-mail: bakulina@gmail.com
** Corresponding author*

Key words: *V-ATPase, a2-subunit, cytohesin-2, ARNO*

Motivation and Aim: V-ATPase is a multiprotein proton pump. It interacts with variety of proteins, including Arf-exchange factor cytohesin-2 (also known as ARNO). Recently it was found that cytosolic part of mouse a2 isoform of a subunit of V-ATPase and six a2-derived peptides interact with ARNO [1]. 3D model of N-terminal cytosolic part of a2-subunit (a2N) is required for understanding details of intermolecular interactions and their functional roles.

Methods and Algorithms: The model is based on the following experimental data: crystal structure of N-terminal part of a homolog of subunit a (PDB ID 3RRK), NMR structures of peptides derived from C-terminal region of a2N, cryo-EM map of *Thermius Thermophilus* V-ATPase [2]. MODELLER and I-TASSER software packages were used for building several protein models, which were fitted into the cryo-EM map with Chimera program, and finally a MODELLER model was selected for the best fit.

Results: The a2N model consists of three domains: the distal lobe (DL), the proximal lobe (PL) and the central bar domain (BD). DL and PL have α/β structures, BD contains two anti-parallel rows of α-helices and links DL and PL. Mapping ARNO interacting peptides shows that there are two distinct sites on a2N, one is formed by two β-strands in DL domain and the loop between them, the other is formed by three β-strands, the adjacent loops and an α-helix in PL domain. We also mapped residues which were determined as contacting with G/E-subunits of V-ATPase by cross-linking experiments with Vph1p, an yeast homolog of a-subunit [3]. These regions occur in close proximity to ARNO interaction sites.

Conclusion: We built the structural model of cytosolic part of mouse a2-subunit of V-ATPase. We identified two ARNO binding sites on distinct domains of a2N. These sites are close, but not overlapping with G/E-subunits binding sites. We hypothesize that binding of ARNO to a2N may modulate its interaction with G/E-subunits of V-ATPase, and could promote disassembly of V-ATPase complex, and, thus, regulate its function.

References:

1. Merkulova, M., Bakulina, A., Thaker, Y. R., Gruber, G., and Marshansky, V. (2010) Specific motifs of the V-ATPase a2-subunit isoform interact with catalytic and regulatory domains of ARNO, *Biochim Biophys Acta,* **1797:** 1398-1409.
2. Lau, W. C., and Rubinstein, J. L. (2012) Subnanometre-resolution structure of the intact Thermus thermophilus H+-driven ATP synthase, *Nature,* **481:** 214-218.
3. Qi, J., and Forgac, M. (2008) Function and subunit interactions of the N-terminal domain of subunit a (Vph1p) of the yeast V-ATPase, *J Biol Chem,* **283:** 19274-19282.

A SIMPLE PERSONAL GENOME VIEWER

Bakulina A.*[1], Diakonov A.[1], Zagrivnaya M.[2]
[1] *State Research Center of Virology and Biotechnology "Vector", Koltsovo, Russia;*
[2] *Pavlov State Medical University of St. Petersburg, St. Petersburg, Russia*
e-mail: bakulina@gmail.com
** Corresponding author*

Motivation and Aim: Genetic diagnosis allows to improve the quality of life. It helps to implement more effective treatments of various diseases and to choose the optimal lifestyle. The role of a genetic testing rises with dramatical cost reducing of personal genome sequencing. However, most of people are not familiar with the capabilities of modern methods of genetic testing and therefore can not benefit from them. Development and distribution of software for personal genome analysis can contribute to solving this problem.

Results: We have developed the OhMyGenes software for storage and unprofessional analysis of user's personal genome. It contains the list of genetic polymorphisms with known effects in the human genome and information where they can be tested, and provides view of variety of the user's genetic test results. The user can view polymorphisms and mutations by category, or perform full-text search of effects or genes descriptions. For some multifactor diseases a post-test risk value is calculated. The database of the software package includes the most important polymorphisms, associated with the risk of monogenic and multifactorial diseases drugs metabolism, athletic status.

Conclusion: Our program helps to solve following problems: promotion of genetic testing among population, informing people about important genetic polymorphisms, storage of testing results in a convinient way.

Availability: The demo-version of the software is available for Windows and Android upon request.

The final version will be available as soon as the database of known polymorphisms, mutations and medical centers is completed enough and the server for automatic update is running.

BIOMARKER CHALLENGE: A CLOUD INSTEAD OF A SET OF THE VANTAGE POINTS

Baranova A.V.
Russian Center of Medical Genetics RAMS, Moscow, Russia,
School of Systems Biology, George Mason University, Fairfax, VA, USA
e-mail: abaranov@gmu.edu

Key words: *biomarkers, prognosis, cancer, whole transcriptome analysis, continuous prognosis model*

Motivation and Aim: To date, the quantification of the diagnostic and prognostic biomarker molecules in the human serum and tissues remains the primary means of enhancing the clinician's ability to predict and detect cancer before it spreads and to predict the outcome of treatment. Importantly, with innumerable molecular markers in development, the discovery of novel standalone markers with acceptable sensitivity and specificity is an extremely rare event. The conventional method to overcome the problem of relatively low sensitivity and specificity of newly discovered biomarkers is to combine them into biomarker panels. However, in many cases these biomarker panels suffer from relatively low reproducibility of results in independently collected sets of samples. This is especially true for the mRNA biomarkers identified by microarray experiments.

Methods and Algorithms: We challenged the biomarker paradigm by developing a distance measure between the entire gene expression profile of a tumor and the center of the space occupied by normal samples. This novel concept allows one to depart from the classical two-bin prediction model (e.g. "bad prognosis/good prognosis") as it produces a continuous prognosis model, where each sample is located in the neighborhood of other samples analyzed post-hoc and associated with known survival.

Results: Whole-transcriptome based distances calculated using Pearson correlation coefficients provide easy visualization of the relative degree of the malignancy characteristic for studied samples. In all studied datasets, on average, tumors were further away from the Normal Sample Space than the paired samples with normal histology. The distance analysis demonstrated remarkable behavioral invariance observed in eighteen independent tumor data sets and provided a robust validation of this approach. The concept of distance analysis is not limited to cancer as it could be generalized to quantify the departure of any given sample from its reference set, i.e. tissue sample of aged persons from reference of non-aged, samples of insulin resistant tissues from normally functioning tissues, and even model cell lines that drift away from the standard phenotype.

Conclusion: If successful, this unconventional approach will shift the tumor biomarker paradigm from expression biomarker panels associated with low reproducibility, to the distance analysis of robust molecular portraits. The proposed distance analysis is versatile in its application as it will be equally attributable to gene expression profiles collected both by microarrays and by RNA-seq platforms.

Availability: on collaborative request.

Acknowledgements: Dr.Ganiraju Manyam (MD Anderson Cancer Center, TX, USA), Dr. Alessandro Jiuliani (Istituto Superiore di Sanita, Roma, Italy), Dr. Boris Veytsman and Lei Wang (George Mason University, Fairfax, USA).

MELANOGENESIS HELPS HUMAN ADIPOSE TISSUE WITHSTAND LOW-GRADE SYSTEMIC IINFLAMMATION

Baranova A.V.
School of Systems Biology, George Mason University, Fairfax, VA, USA
e-mail: abaranov@gmu.edu

Purpose: To curtail secondary morbidities in obesity and overweight individuals without the need of the weight loss.

Methods: Molecular pathways specifically activated in adipose of morbidly obese individuals were studied by microarray and by various biochemical assays.

Results: The melanin, common skin pigment, was discovered in human adipose tissue. A marked heterogeneity of melanin content was observed in individual adipose tissue extracts. Melanogenesis was shown to be excessively stimulated in morbid obesity. Positive correlation between fasting glucose levels and total outputs of the melanogenic pathway in adipose tissues was observed. Novel hypothesis stating that ectopic synthesis of melanin may serve as a compensatory mechanism that utilizes its anti-inflammatory and its oxidative damage absorbing properties was developed. With the progression of obesity and an increase of the cellular fat deposition, adipocytes become more exposed to endogenous apoptotic signals, especially ROS. To counteract pro-apoptotic ROS effects, the adipocytes in turn may ectopically activate the genetic program of melanogenesis, thus neutralizing excessive ROS. Adipocytic melanin would also suppress the secretion of pro-inflammatory molecules, thereby decreasing the pro-inflammatory background in obese subjects and alleviating the metabolic syndrome. High polymorphisms of human genes regulating melanin biosynthesis may account for the differences in propensity to develop secondary complications of obesity.

Conclusion: Molecular compounds stimulating melanogenesis, particularly, the synthetic agonists of α-MSH receptors, have already been proven safe in human trials for therapeutic tanning. These compounds shall be tested as the preventive medications aimed at curtailing the development of devastating metabolic complications in obese and overweight populations.

BIOINFORMATICS IN TRANSLATIONAL RESEARCH

Baranova A.V., Chandhoke V.
School of Systems Biology, George Mason University, Fairfax, VA, USA
e-mail: abaranov@gmu.edu, vchandho@gmu.edu

Key words: *biomarkers, prognosis, cancer, whole transcriptome analysis, continuous prognosis model*

Translational bioinformatics hold a number of promises, namely, an ability to derive pathogenic mechanism from the gene interactions, to ascertain effectiveness of prevention strategy for a given genotype, compare effectiveness of treatments, infer patient-specific drug doses and the possibilities for the development of given side effect. One day, these promises may amount to a truly "tailored", individualized treatment. However bright this perspective might be, it is important to remember that clinical data analysis routinely encounters various data biases that substantially influence the accuracy of the derived conclusions. In this presentation, we will sort through typical biases present in the clinical datasets and wil use specific examples how to overcome these hidden perils.

INBREEDING AND DIFFERENTLY DIRECTED DYNAMICS OF ISSR-PCR AND IRAP-PCR MARKERS POLYMORPHISM IN MUSK OXEN POPULATIONS

Barducov N.V.*, Sipko T.P., Glazko V.I.
Russian State Agrarian University–MTAA, Moscow, Russia;
A.N. Severtsov Institute of Ecology and Evolution of RAS, Moscow, Russia
e-mail: bardukv-nikolajj@mail.ru
** Corresponding author*

Key words: *population genetic structure, inbreeding, polymorphism, ISSR-PCR markers, IRAP-PCR markers*

Motivation and aim: For the control of genetic structure of three musk oxen (*Ovibos moschatus*) populations a comparative analysis of gene pools of these populations by means of ISSR-PCR (Inter-Simple Sequence Repeat) and IRAP-PCR (Inter-Retrotransposon Amplified Polymorphism) was carried out. In populations which differ by inbreeding coefficient (F) polymorphism estimation vary depending on type of markers used: increasing F was followed by decrease in ISSR-PCR but by increase in IRAP-PCR markers.

Methods and Algorithms: An investigation is done on 80 musk oxen from three populations – source in Eastern Greenland (1) and introduced into Wrangel island (2) and Taymyr Peninsula (3). For multiloci scanning we used two PCR methods: ISSR-PCR and IRAP-PCR. In the former we used $(AG)_9C$, $(GA)_9C$ and $(GAG)_6C$ as primers, while in the latter terminal flanks of retrotransposons LTR-SIRE-1 и PawS 5 were taken. Amplification spectra containing 57 DNA fragments of different length in summary were achieved in which every amplicon was considered as a single locus. A PIC index (Polymorphic Information Content) was counted for each amplicon and also an average value for each primer. Shares of polymorphic loci were counted for three musk oxen populations separately. According to M. Nei (DN, 1972) genetic distances were estimated which were used to make a dendrogram by using TFPGA software.

Results: In ISSR-PCR spectra we found population specific frequencies of occurrence of DNA fragments of different length flanked by inverted repeats of microsatellites. In summary, the highest PIC values were observed in source population from Greenland (average PIC = 0.1) then Taymyr population originated from animals from islands Banks and Nunivak (PIC = 0.08) and the lowest were observed in Wrangel population (PIC = 0.06) which has the largest inbreeding degree. So, ISSR markers showed the lowest PIC values with the highest F. A dendrogram built on DN counted from ISSR data corresponded the history of populations' origin. IRAP-PCR markers showed an opposite trend for PIC index: Greenland population had an average PIC = 0.02, Taymyr – 0.13 and Wrangel – 0.14. Thus, the highest PIC counted from IRAP data is observed in Wrangel population which has the largest inbreeding degree.

Conclusion: By using of two types of molecular markers we found differences in estimation of genetic structure of musk oxen populations investigated: population inbreeding history coincided with decrease of polymorphism by ISSR markers but differed from observed level of polymorphism by IRAP markers. The data acquired evidence that multiloci genome scanning may reflect differences in gene pool processes of animals depending on genotyped elements: increase of homozygosity under inbreeding counted by inverted repeats of microsatellites and its decrease counted by terminal flanks of retrotransposons.

HIGH-THROUGHPUT SCREENING FOR THE DEVELOPMENT OF NOVEL SELECTIVE LIGANDS OF D_2 DOPAMINE RECEPTORS

Barnaeva E.*[1], Free R.B.[2], Hu X.[1], Southall N.[1], Bryant-Genevier M.[1], Titus S.[1], Ferrer M.[1], Marugan J.[1], Sibley D.R. [2]

[1] *National Center for Advancing Translational Sciences;*

[2] *National Institute of Neurological Disorders and Stroke, National Institutes of Health, 9800 Medical Center Drive, Rockville, Maryland 20850, United States*

e-mail: barnaevaes@mail.nih.gov

Key words: *high-throughput screening (HTS), D_2 dopamine receptor, functional selectivity*

The G-protein couple receptor (GPCR) D_2 dopamine receptor (D2 DAR) is an important therapeutic target for the treatment of a number of neuropsychiatric disorders, including Parkinson's disease. As most drugs targeting the D2 DAR are non-selective, there is great interest in the development of novel selective ligands of D2 DAR to further clarify the pharmacological role of this receptor for the treatment of different diseases. A novel approach for attaining greater selectivity of drugs targeting GPCRs is to identify small molecule ligands that exhibit "functional-selectivity". The phenomenon of "functional selectivity" can occur when activation of a GPCR transduces signals through different intracellular pathways, such as the traditional G-protein and second messengers (cAMP and Ca^{2+}), or β-arrestin and pERK. It has been shown that for some GPCRs efficacy and toxicity effects of ligands might be driven by activation of different signaling pathways, and compounds that are functionally selective might provide a better therapeutic window in clinic.

We recently have completed high throughput-screening (HTS) of 400,000+ small molecules in a variety of D2 DAR assays to identify functionally selective agonists and antagonists of this receptor. One assay measured D2 DAR signaling through Ca^{2+} by using a fluorescent calcium flux assay; the second assay measured signaling through β-arrestin by using a D2 DAR β-arrestin Enzyme Complementation cell line and PathHunter Assay from DiscoveRx. Both assays were miniaturized to 1536-well format and shown to be robust for automated HTS.

Data will be presented demonstrating the quality of screens implemented and analysis of the results focusing on selectivity between the two assays.

SMALL NON-CODING RNAs OF HUMAN BLOOD PLASMA OF HEALTHY DONORS AND PATIENTS WITH NON-SMALL CELL LUNG CANCER

Baryakin D.N.*[1], Semenov D.V. *, Brenner E.V. [1], Kurilshikov A.M. [1], Kozlov V.V.[2], Narov Y.E.[2], Vasiliev G.V.[3], Bryzgalov L.O.[3], Chikova E.D.[1], Filippova J.A.[1], Kuligina E.V. [1], Richter V.A.[1]

[1] Institute of Chemical Biology and Fundamental Medicine SB RAS, Novosibirsk, Russia;
[2] Novosibirsk Regional Cancer Centre, Novosibirsk, Russia;
[3] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
** Corresponding author*
e-mail: baryakindn@niboch.nsc.ru

Key words: *high-throughput sequencing, circulating RNAs, miRNA, non-small cell lung cancer*

Motivation and Aim: Circulating nucleic acids are subject of numerous modern researches aimed at establishing of new mechanism for distant regulation of physiological processes, and at using of extracellular nucleic acids as diagnostic and prognostic markers of pathological processes in the organism.

The aim of this study is a detailed description of structure and forms of short extracellular RNAs of human plasma, aimed at identifying new forms of regulatory RNA and the establishment of the mechanism of their action, and the development of unique and complex diagnostic markers of human diseases.

Methods and Algorithms: In this study we analyze the diversity of short non-coding RNA blood plasma of 8 healthy volunteers and 8 patients with non-small cell lung cancer. In order to obtain cDNA libraries that encode the most full-scale set of circulating RNAs, short (n> 19) plasma RNAs were exposed to dephosphorylation followed by 5'-phosphorylation, ligation with adapters, reverse transcription and amplification. Individual cDNA libraries were sequenced on a platform SOLiD (V.3). The resulting data sets were analyzed using the Bowtie/Cufflinks software.

Results: It was found that human blood plasma contains fragments of rRNA, tRNA, mRNA transcripts of the mitochondria, the mature miRNA, scRNA, snRNA, snoRNA, as well as fragments of transcripts not annotated previously.

A detailed analysis of miRNAs and miRNA-like forms including determination of the most represented form of known human miRNAs in the blood plasma, finding new miRNA-like forms using the software mirDeep 2.0; identifying potential mRNA target of circulating miRNAs and candidate miRNAs.

Conclusion: Comparative analysis of plasma miRNA of healthy donors and patients with non-small cell lung cancer allowed us to characterize a set of changes in expression profile of extracellular human microRNAs at the origin and development of malignant tumors.

Acknowledgements The work was supported by the Russian Ministry of Science and Education (02.740.11.0715); RFBR grants № 10-04-01442-a; № 10-04-01386-a.

MODELLING DRUG RESISTANCE IN BREAST CANCER THROUGH NETWORK RECONSTRUCTION BASED ON LONGITUDINAL PROTEIN ARRAY DATA

Beissbarth T.
University of Goettingen, Goettingen, Germany
e-mail: Tim.Beissbarth@med.uni-goettingen.de

Motivation and Aim: Network inference from high-throughput data is an important means for the analysis of biological systems. For instance, in cancer research, the functional relationships of cancer related proteins, summarised into signalling networks are of central interest for the identification of pathways that influence tumour development. *De novo* reconstruction of signalling pathways from data allows to unravel interactions between proteins and make qualitative statements on possible aberrations of the cellular regulatory program.

Results: We developed the method Dynamic Deterministic Effects Propagation Networks (DDEPN) for reconstructing signalling networks from time course experiments after external perturbation and show an application of the method to data measuring abundance of phosphorylated proteins in a human breast cancer cell lines, generated on reverse phase protein arrays. This modeling approach is a special type of Bayesian Network. Signalling dynamics is modelled using active and passive states for each protein at each timepoint. A fixed signal propagation scheme generates a set of possible state transitions on a discrete timescale for a given network hypothesis, reducing the number of theoretically reachable states. Breast cancer cell lines are used as model system to study the cellular response to drug treatments in a time-resolved way.

Conclusion: Based on these kind of data, our modelling approach allows to generate hypotheses on molecular interactions in the signalling cascades from the ERBB signalling pathway and helps to understand drug resistance mechanisms in different subtypes of breast cancer.

SEQUENCING AND *DE NOVO* TRANSCRIPTOME ASSEMBLY OF *STELLARIA MEDIA* (L.) VILL.

Belenikin M.S.[1, 2], Speranskaya A.S.*[1, 3], Melnikova N.V.[1], Oparina N.Y. [1], Darii M.V.[1, 4], Dmitriev A.A. [1], Slavokhotova A.A.[5], Korostyleva T.V.[5], Kudryavtseva A.V. [1], Odintsova T.I. [5]

[1] Engelgardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow, Russia;
[2] Research Institute of Physico-Chemical Medicine of Russian Federal Medico-Biological Agency, Moscow, Russia;
[3] Lomonosov Moscow State University, Department of Biology, Moscow, Russia, e-mail: hannadt@mail.ru;
[4] Lomonosov Moscow State University, Department of Chemistry, Moscow, Russia;
[5] Vavilov Institute of General Genetics, Russian Academy of Sciences, Moscow, Russia
** Corresponding author*

Key words: *Next-generation sequencing, RNA-seq, Illumina, Stellaria media, de-novo transcriptome assembly, Trinity, Oases*

Motivation and Aim: Common chickweed Stellaria media (Caryophyllaceae) is cool-season annual weed native to Europe and Russia. This plant is used in traditional medicine and shows effective antiviral activities. A number of novel antimicrobial peptides, components of innate immunity were isolated from this plant.

Methods and Algorithms: Transcriptome sequencing of 5-old-day seedlings of Stellaria media grown under natural conditions (25oC day and 18oC night temperature) was performed using next-generation sequencing platform Illumina Genome Analyser and 51x106 of pair-end raw 76 bp sequences were generated accordingly. The transcriptomes were assembled using single k-mer assemblers Oases (v.0.2.01) и Trinity (v.2012-03-17).

Results: We have performed de novo assembly of obtained paired-end reads using different single k-mer assembly approaches. The higher stringency options were applied using Oases assembler with 31 k-mer value. This method allowed us to produce ~40,000 transcripts longer than 300 bp. The longest contig was of 5 Kbp length. The total length of assembled transcriptome was equal to 22 Mbp. We have carried out de novo assembly with Trinity method [3]. Using the same threshold of 300 bp for assembled transcripts, we have produced more than 100,000 transcripts (total transcriptome length was equal to 112 Mbp). The length of the longest contig was 9 Kbp. Almost all Oases-assembled contigs were homologous to Trinity-assembled.

Conclusion: We have constructed and annotated the consensus transcriptome assembly of Stellaria media vegetative tissues. The combined usage of two different assemblers proved itself as a reliable strategy. We present for the first time the largest dataset of cDNAs (~100000 sequences) for this plant and also for the total Caryophyllaceae family (characterized with total 112,163 mRNA sequences in NCBI Genbank).

This work was supported by grant №16.552.11.7034 of the Russian Ministry of Education and Science.

STUDY OF INTERINDIVIDUAL VARIABILITY OF WARFARIN DOSAGE AMONG POPULATION OF THE WESTERN SIBERIAN REGION OF RUSSIA

Belozertseva L.A.*, Voronina E.N., Koh N.V., Cvetovskaya G.A., Lifshits G.I., Filipenko M.L.
Institute of Chemical Biology and Fundamental Medicine SB RAS, Novosibirsk, Russia
e-mail: white_lilia89@mail.ru
** Corresponding author*

Key words: *warfarin, VKORC, CYP2C9*

Motivation and Aim Warfarin is an oral anticoagulant widely used around the world. The molecular target of warfarin is enzyme vitamin K epoxide reductase (VKOR), which reduces the oxidized form of vitamin K to hydroquinone. The reduced vitamin K is cofactor of gamma-glutamyl carboxylase (GGCX), which provides the carboxylation and, thereby, activates coagulation factors II, VII, IX, and X and proteins C, S, and Z. Warfarin inhibits VKOR that slows hemostasis. In the appointment of warfarin, patients show a wide range of interindividual differences in drug doses needed to achieve the desired therapeutic effect. Thus, the aim of our study was to investigate the influence of polymorphic variants of *VKORC1* (-1639 G> A, rs9923231), *CYP2C9* (* 2,* 3), *GGCX* (12970 C> G, rs11676382), *PROC* (2583 A> T, rs1799810), *FVII* (10976 G> A, rs6046), *CYP4F2* (23454 G> A, rs2108622) genes, and non-genetic parameters on the variability of warfarin dose among the population of the West Siberian region of Russia.

Methods and Algorithms The group of patients taking warfarin (N=113) contain patients treated in the CNMT in Novosibirsk. DNA was extracted from venous blood using standard procedures. Determination of genotypes of polymorphic loci was performed by Real-time PCR. Statistical analysis was performed using the program Statistica for Windows, v.8 (StatSoft, Inc.).

Results Warfarin dose significantly varied between carries of different genotypes polymorphic loci of *VKORC1* (-1639 G>A) and *CYP2C9* *3 (p=0,00001 и p=0,018 respectively). The media dose amounted 6,25 mg/day for G/G-carries (N=47), 4,4 mg/day for G/A-carries (N=57), 2,5 mg/day for A/A-carries (N=9) of *VKORC1* -1639 G>A; 5 mg/day for *1/*1-carries (N=101), 3,75 mg/day for *1/*3-carries (N=12). According linear regression results the -1639 G>A *VKORC1* and *CYP2C9*3* variants accounted for 20,7 and 5,7% of the warfarin dosage variability in the patients studied (p=0,000001 и 0,00629 respectively). None of the other polymorphic loci was statistically significantly associated with dose (*CYP2C9*2* (p=0,3579), *GGCX* 12970 C>G (p=0,3745), *PROC* 2583 A>T (p=0,7262), *FVII* 10976 G>A (p=0,4045), *CYP4F2* 23454 G>A (p=0,2844)). According to multiple linear regression results, -1639 G>A *VKORC1* and *CYP2C9*3* together could explain about 22,3 % of the total warfarin dose variation in our group of patients.

Conclusion According to our results, polymorphic variant -1639 G>A in *VKORC1* gene can explain 20,7 %, allele *CYP2C9*3* - 5,7 % of warfarin dose variation among the population of Western Siberian.

IDENTIFICATION OF STEM CELL GENES IN THE FLATWORM *MACROSTOMUM LIGNANO*

Berezikov E.*, Simanov D., Mouton S., Arindrarto W., Van Nies K., de Mulder K.
Hubrecht Institute and University Medical Center Utrecht, Utrecht, The Netherlands
e-mail: e.berezikov@hubrecht.eu
** Corresponding author*

Motivation and Aim: Macrostomum lignano is a free-living flatworm with high regeneration capacity facilitated by stem cells called neoblasts [1-3]. Due to its high regeneration capacity, small size, transparency and clear morphology, ease of culture, short generation time and amenability to genetic manipulation, *M. lignano* has great potential as a model organism for stem cell research. Yet due to the novelty of this model organism little is known about genes involved in regulation of stem cells in *M. lignano*. As the first step towards elucidation of gene regulatory programs that control neoblast biology, we aim to identify neoblast markers – genes that are specifically expressed in stem cells of *M. lignano*.

Methods and Algorithms: In order to characterize the transcriptome of *M. ligano* we have generated RNA-Seq data using 454 and Illumina platforms and performed *de novo* transcriptome assembly. Next, we used comparative RNA sequencing from several developmental stages and from irradiated animals (which are depleted in neoblast) to generate a list of candidate neoblast genes. Expression patterns of the candidates were investigated by whole mount *in situ* hybridization, and functional roles of genes with clear expression in neoblast were studied by RNAi.

Results and Conclusions: Using stringent cut-off criteria we have identified 170 genes differentially expressed between irradiated and non-irradiated animals and with clear homologs in human. These genes included several known neoblast markers (such as *piwi* and *pcna*), many genes with a known role in cell cycle regulation, as well as numerous genes with previously unknown function. The majority of the candidates appeared to have neoblast- or germline-specific expression patterns. For several genes without previously known function, knockdown by RNAi resulted in loss of tissue homeostasis and regeneration failure, suggesting essential roles of these genes in neoblast regulation.

References:

1. Ladurner et al. (2005). A new model organism among the lower Bilateria and the use of digital microscopy in taxonomy of meiobenthic Platyhelminthes:Macrostomum lignano, n. sp. (Rhabditophora, Macrostomorpha). *JZS* **43:** 114-126.
2. Pfister et al. (2008). Flatworm stem cells and the germ line: developmental and evolutionary implications of macvasa expression in Macrostomum lignano. *Dev. Biol*. **319**:146-159.
3. De Mulder et al. (2009). Stem cells are differentially regulated during development, regeneration and homeostasis in flatworms. *Dev Biol* **334**:198-212.

DISCOVERING THE EPIGENOME: GLOBAL MAPPING OF HISTONE MARKS AND MODELING TRANSCRIPTIONAL MEMORY

Binder H.*, Galle J., Rohlf T., Prohaska S., Hopp L., Steiner L., Wirth H.
Interdisciplinary Centre for Bioinformatics, University Leipzig, Germany
e-mail: binder@izbi.uni-leipzig.de
** Corresponding author*

Key words: *histone methylation, gene activity, transcriptional memory, machine learning, theoretical model*

Motivation and aim: Epigenetic mechanisms play an important role in regulating and stabilizing functional states of living cells. However, in spite of an increasing amount of experimental data, genome-wide mapping and visualization methods as well as models of transcriptional regulation by epigenetic processes are rather rare. In this review, we focus on epigenetic modes of transcriptional regulation based on histone modifications and their potential dynamical interplay with DNA methylation and higher-order chromatin structure. We discovered the epigenome and its relation to the transcriptional activity of the affected genes in a two-way study combining a top-down data-driven analysis of genome-wide histone modifications and a down-top modeling approach based on first principles given as basal histone-modification reactions.

Data analysis: The data-analysis makes use of genome-wide ChIP-seq data on different histone marks. We present a novel method for annotation-independent exploration of epigenetic data and their inter-correlation with other genome-wide features such as gene expression and DNA-methylation patterns. The approach allows lossless compression of the information about epigenetic states and is followed by sorting, clustering and visualization via self-organizing maps. Furthermore, complementary quantitative features obtained from the genomic sequence, localization or gene expression can be explored to detect possible correlations with the modification states. The method actually provides a global view of genome-wide epigenetic information. It allows tracing formation and disappearance of combinations of different histone modifications over different cell samples.

Theoretical Modeling: Our in-silico modeling addresses transcriptional regulation in cells based on chromatin-related mechanisms to ensure that functional states can be maintained and adapted to variable environments. Our model of transcriptional regulation describes binding of protein complexes to chromatin which are capable of reading and writing histone marks. Molecular interactions between these complexes, DNA and the histones create a regulatory switch of transcriptional activity possessing a regulatory memory. The regulatory states of the switch depend on the activity of histone (de-) methylases, the structure of the DNA-binding regions of the complexes, and the number of histones contributing to binding. We apply our model to transcriptional regulation by trithorax- and polycomb- complex binding.

Finally, results of data analysis and modeling are combined: Histone modification data on pluripotent and lineage-committed cells are used to validate the basic model assumptions and to provide evidence for cooperative histone modifications within extended chromatin regions. Our results provide new insights into epigenetic modes of transcriptional regulation and represent a basic step towards multi-scale models of this process.

References:

1. Binder H, Steiner L, Wirth H, Rohlf T, S P, Galle J: Transcriptional memory emerges from cooperative histone modifications http://precedingsnaturecom/documents/6507/version/1 2011, preprint.
2. Rohlf T, Steiner L, Przybilla J, Prohaska S, Binder H, Galle J: Modeling the Dynamic Epigenome: from histone modifications towards self-organizing chromatin. Epigenomics 2012, in press. (see http://www.izbi.uni-leipzig.de/izbi/mitarbeiter/Binder/Rohlf.pdf)

THE LOCATION OF T1 DIABETES ASSOCIATED SNPs IN REGULATORY REGIONS

te Boekhorst R.*[1], Beka S.[1], Abnizova I.I.[2]
[1] *School of Computer Science, University of Hertfordshire, Hatfield, UK*
[2] *Wellcome Trust Sanger Institute, Hinxton, UK*
e-mail:r.teboekhorst@herts.ac.uk
** Corresponding author*

Key words: *T1 diabetes SNP regulation*

Motivation and Aim: Although many association studies on complex diseases focus on variation in coding DNA, recent research shows increasing evidence that the cause of such diseases should be sought in the regulation of gene activity. Rather than studying mutations in genes coding for transcription factors, our work focuses on genetic variants (SNPs) in regulatory modules (TFBS, enhancers, promoters, or other genic locations likely to be involved in gene regulation such as UTR, introns and splice junctions) that are in Type 1 Diabetes (T1D) susceptibility regions of the human genome. The specific research question is: are SNPs associated with T1D more likely to occur in (putative) regulatory regions than other SNPs found in T1D susceptible regions? In addition: Because genes may overlap and/or occurrence in multiple transcripts, one and the same SNP may be associated with more than one genic location and affect more than one functional region (e.g. is a mutation in as well a coding region as a regulatory region). Are SNPS associated with a complex disease such as T1D more likely to be of this kind?

Methods: An extensive search in the databases ENSEMBL and T1Dbase was conducted to collect information on all SNPs in T1D susceptible regions [coordinates on genome, type of variant (mutant allele, wildtype allele, transversion or transition, synonymous or non-synonymous), transcriptID, intra-genic region (up/downstream, exon, intron, splice junction, UTR, type of (micro)RNA), type of gene (coding/pseudo/microRNA) and associated diseases].

A statistical analysis was performed to assess possible associations between the status of a SNP (associated/not-associated with T1D) on the one hand and features characterising the (intra/inter-) genic position on the other hand.

Results: 1) SNPs associated with T1D are more likely to occur in regulatory regions than those in the same susceptible region but that that are not associated with T1D

2) SNPs associated with T1D occur more often in multiple genic locations than those that occur in only one genic location. They appear to be relatively over-abundant in transcription factor binding sites, introns and splice-sites.

Conclusion: The results show the importance of mutations in regulatory regions in the occurrence of type 1 diabetes.

Availability: The data are available from the authors.

FLUORESCENCE *IN SITU* HYBRIDIZATION WITH CHROMOSOME-DERIVED DNA PROBES ON *OPISTHORCHIS FELINEUS* AND *METORCHIS XANTHOSOMUS* CHROMOSOMES WITHOUT SUPPRESSION OF REPETITIVE DNA SEQUENCES

Bogomolov A.G.*[1], Zadesenets K.S.[1], KaramyshevaT.V. [1], Podkolodnyy N.L.[1, 2], Rubtsov N.B.[1, 3]

[1] *Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia, e-mail: mantis_anton@bionet.nsc.ru;*
[2] *Institute of Computational Mathematics and Mathematical Geophysics SB RAS, Novosibirsk, Russia;*
[3] *Novosibirsk State University, Novosibirsk, Russia*
** Corresponding author*

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

Motivation and Aim: For chromosome painting with the probes derived from individual chromosome suppression of repetitive DNA sequences [1] is standard used. However in some cases, suppression can't be done – DNA of some species is not available in required amount. This paper presents a method that allows us to improve the results of chromosome-derived DNA probe hybridization performed without suppression of repetitive DNA sequences hybridization.

Methods and Algorithms: The method treated the images of two-color FISH with chromosome-derived microdissected DNA probes [2]. The input images stored results of hybridization with DNA probes in different channels. Also there was a channel including the image of DAPI staining. The total signal at the point of grayscale image (this grayscale image is the intensity distribution of the signal from one fluorochrome), which contained the results of hybridization with DNA probe, could be represented as a sum of signals from the chromosome-specific sequences (specific signals) and non-specific signals (including signal from dispersed repeated sequences and noise). Chromosomes, which DNA probes were derived from, showed both types of signal. On another channel, the chromosomes showed only non-specific signal. This was due to the fact that the DNA probe was derived from another chromosome. The difference of signals in the channels was due to the registration of specific signal in the first channel. But the resulting intensity depended on many factors. We have to carried out make the following procedures with the images: filtering, semi-automatic threshold segmentation, labeling of objects, pattern recognition, normalization of signals and calculation of specific signal, linear staining of the result image.

In result the chromosomes could be divided into three classes: chromosomes, which one of the DNA probes were derived from, and chromosomes containing only non-specific signal. We use correlation of signals from two channels as an objects feature for pattern recognition. The Irvins criterion [2] and the consequence of the theorem of compactness classes in feature space are used to determine the number of classes in the image.

Results: The method was tested on images of human metaphase chromosomes and meiotic chromosomes of *O. felineus* and *M. xanthosomus*. The obtained results showed, that the method in general allowed to identify specific FISH signal without suppression of repetitive DNA hybridization. In addition it allowed to detect specificity of the distribution of different DNA repeated sequences types.

References:

1. Lichter P., Cremer T., Borden J., Manuelidis L., Ward DC. (1988) Delineation of individual human chromosomes in metaphase and interphase cells by in situ suppression hybridization using recombinant DNA libraries. // Human genetics. V. 80, N. 3. P. 224-234
2. Zadesenets KS, Karamysheva TV, Katokhin AV, Mordvinov VA, Rubtsov NB Distribution of repetitive DNA sequences in chromosomes of five opisthorchid species(Trematoda, Opisthorchiidae) // Parasitol Int (2012), V.61, P. 84-86.
3. Tretiak L.N. (2004) Processing results of observation. Orenburg: OSU, P. 44-45

RANDTRAN: RANDOM TRANSCRIPTOME SEQUENCE GENERATOR

Borzov E.A.* [1], Marakhonov A.V. [1], Baranova A.V.[1, 2], Skoblov M.Yu.[1, 3]
[1] *Federal State Budgetary Institution "Research Centre for Medical Genetics" under the Russian Academy of Medical Sciences, Moscow, Russia;*
[2] *School of Systems Biology, College of Science, George Mason University, Fairfax, VA USA;*
[3] *State Budgetary Institution of Higher Education "Moscow State Medical and Dental University", Moscow, Russia*
e-mail: eborzov@generesearch.ru
** Corresponding author*

Key words: *random sequence, transcriptome, generator*

Motivation and Aim: Random transcriptome generation is a frequently used method in modern bioinformatics researches. However algorithms of its creation are not widely discussed in the literature and the concept of making of random transcriptome is often limited to conserve mononucleotides frequency. Our aim was to make the program for random transcriptome generation that takes into account the structure of transcript (5'-UTR, CDS, 3'-UTR, or ncRNA) with its own dinucleotide and trinucleotide (codon) frequencies for different species.

Methods and Algorithms: We have used classic Monte-Carlo methods. The program is written in Perl using Tk Module for interface (http://www.cpan.org/).

Results: The program «RANDTRAN» has user-friendly interface and flexible options to create your own random transcriptome. It used a standalone .txt frequency file (2KB) that defined a lot of parameters for specific organism. For the moment the program is packaged with 23 files (for 11 eukaryotes and 12 bacteria and archea species). The frequency file contains a lot of parameters described special features, such as dinucleotide and trinucleotide frequencies for each transcript region, length distribution data of each part of transcript, quantity of coding and non-coding transcripts. The advantage of our program is to allow changing of above arguments or creating new frequency file that gives a user the opportunity to generate the transcriptome with necessary characteristics. The detailed manual is attached.

Availability: Download from our website http://www.generesearch.ru/research.html

MASSIVE PARALLEL EXON SEQUENCING AS FUNDAMENTAL APPROACH IN STUDYING SNPs THAT CAN LEAD TO ALZHEIMER DISEASES

Boulygina E.S.[1], Nedoluzhko A.V.[1], Tsygankova S.V.[1], Tchekanov N.N.[2], Mazur A.M.[2], Artemov A.V.[3], Prokhortchouk E.B.[1, 2], Skryabin K.G.[1, 2]

[1] National Research Center "Kurchatov Institute", Kurchatov Sq 1, Moscow, Russia;

[2] Centre "Bioengineering" of the Russian Academy of Sciences, Moscow, Russia;

[3] T-gene LTD, Dubna, Moscow region, Russia

The members of Russian family that had Alzheimer disease in three generations were used to study genetic variations that are caused of this disease. Analysis of 249 Alzheimer associated genes was performed using SureSelect approach capturing the coding regions of the genes (including 200 bp intron sites), 5' promoter regions (up to 1000 bp in 5' region from the start point of gene transcription) and 3' noncoding regions (up to 1000 bp in 3' region from a site of polyadenilation of mRNA). Trapped DNA was sequenced with SOLiD 4.0 («Applied Biosystems», the USA). The analysis of the received data has shown, that 80 % of all short reads mapped on the chosen sites of human genome (hg18) with average depth of a coverage 313. SNP calling revealed 2335 SNPs (663 (31 %) within exons) and 166 short inserts/deletions (33 within exons) in analyzed sites. Neither nonsynonymous SNP's nor short indels were found within crucial Alzheimer associated genes (APP, PSEN1, PSEN2 and APOE).

Using the software developed by us the polymorphisms which are characteristic for all family N members with diagnosed Alzheimer's disease have been selected. Also the absence of contradictions in the found genotypes between parents and their children was checked. Taking into account the above described restrictions only those polymorphisms which are in coding regions of genes and lead to amino acid replacement (missense) or to occurrence of a stop codon in a gene (nonsense) have been selected for the further analysis. Then each of selected has been checked up using SIFT and ANNOVAR programs to evaluate its influence on corresponding protein sequence. As a result only 16 SNPs, having the maximum probability of association with Alzheimer's disease have been chosen for the subsequent analysis. The confirmation of those SNPs were done by Senger sequencing on ABI 3730. The distribution of these SNPs within Russian population is currently studied.

Potential variants which can lead to development of this disease are revealed, and some of these variants are new and were not described previously. Information on variants of genes will allow to learn more about molecular etiology of disease, and also to use simple methods of diagnostics (PCR) for personal predisposition to the disease.

HIGH-THROUGHPUT SEQUENCING OF MYCOBACTERIUM STRAINS USED FOR STEROID COMPOUNDS BIOSYNTHESIS

Bragin E.Yu.*[1], Ashapkin V.V., Shtratnikova V.Yu.[1], Schelkunov M.I., Dovbnya D.V., Donova M.V.

[1] Innovative Technology Center "Biologically Active Compounds and Application", Russian Academy of Sciences, Moscow;

[2] G.K.Skryabin Institute of Biochemistry & Physiology of Microorganisms, Russian Academy of Sciences, Pushchino, Moscow Region

e-mail: bragory@yandex.ru

** Corresponding author*

Key words: *steroid bioconversion, Mycobacterium, whole-genome sequencing*

Motivation and Aim: Strains of *Mycobacterium* spp VKM Ac-1815D, 1816D, 1817D are used for bioconversion of phytosterol to androst-4-en-3,17-dione (AD), androsta-1,4-diene-3,17-dione (ADD) and 9a-hydroxy androst-4-ene-3,17-dione (9-OH-AD) respectively. Aim of this study - investigation of genes and operons of steroid catabolism of these strains.

Methods and Algorithm: High-throughput sequencing was performed with Genome Analyser IIx (Illumina), pair-end reads, 72+72 b.p. Genomes assembling was performed with Velvet software. Search of genes was made by Blast (NCBI) and NCBI Genome Workbench software. Operons was determined with internet-service FgenesB.

Results: We find out 36-38 transcriptional units associated with steroid catabolism in strains VKM Ac-1815D and VKM Ac-1816D, and 64 in VKM Ac-1817D. The key role in steroid catabolism (start of destruction of steroid rings) play 3-ketosteroid dehydrogenases (kstD) and 3-ketosteroid-9-alpha-hydroxylases (consist of kshA and kshB). A single gene kstD is in VKM Ac-1815D and in VKM Ac-1816D, a one SNP between kstD of these strains probably inactivate this gene in VKM Ac-1815D. 4 genes kstD are in VKM Ac-1817D, although kstD activity in this strain expected blocked. 2 genes kshA and 1 gene kshB are in strains VKM Ac-1815D and VKM Ac-1816D and no mutations between them. Strain VKM Ac-1817D have 5 genes kshA and 2 genes kshB, which can define high accumulation of 9-OH-AD. So biochemical characteristics of this strain should depend on activity of another genes or regulated regions.

Conclusion: We obtain long genome sequences of strains VKM Ac-1815D, 1816D, 1817D, and this will allow to investigate SNP and regulatory regions for understanding of steroid catabolism in these strains. We determine genes and operones included in the steroid catabolism of strains VKM Ac-1815D, 1816D, 1817D. It allows to select targets for improvement of biotecnological characteristics of these strains.

APPLICATION OF CONFORMATIONAL PEPTIDES FOR ANALYSIS OF ALLERGENIC PROTEINS

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

Key words: *allergy, allergen, conformational peptides*

Motivation and Aim: Nowadays more than one third of the world population is affected from various allergic diseases. One of the most effective ways to prevent development of allergy is the elimination therapy. For this reason, development of methods for assessing allergenic properties of proteins is a major challenge. Currently existing *in silico methods of assessment* for protein allergenicity use only information about the amino acid sequences of proteins. Information about three-dimensional structure of proteins by such methods usually is not taken into account.

Methods and Algorithms: We have developed a method for representing a surface of protein molecule as a set of short amino acid sequences. These short sequences were called conformational peptides. Conformational peptides were calculated according to the following rules: 1) Two amino acids were accepted as bound in a conformational peptide, if the distance between their C-alpha atoms in protein 3D structure was not greater than 5E. 2) The average residue solvent accessibility for amino acids for a conformational peptide was not less than 50%. 3) The length of the conformational peptides and the linear peptides were the same (8 amino acid residues).

Allergenicity was predicted through search of such peptides from allergenic proteins in query protein.

Results, Conclusion and Availability: We have developed a method for allergenicity prediction, which uses information about the conformational peptides. This method is able to predict the allergenicity using only information about the primary sequence of a query protein. The allergenicity prediction method was integrated into the Protein Structure Discovery System (http://www-bionet.sscc.ru/psd/cgi-bin/programs/Allergen/allergen.cgi). A database of conformational peptides, calculated for three-dimensional structures of allergenic proteins, was created. It was shown that improvement of the allergenicity prediction method can be achieved by using conformational peptides.

Acknowledgements: work is supported by Russian Ministry of Education and Science, contract No. 07.514.11.4003.

PARALLEL NETWORK ANALYSIS ON INTEGRATED LIFE SCIENCE DATA

Braun D.
Bio-/Medical Informatics Department, Bielefeld, Germany
e-mail: dbraun@techfak.uni-bielefeld.de

Key words: *Parallel Computing, Data Integration, Network*

Motivation and Aim: One main issue in bioinformatics is the representation and organization of multidimensional high volume data. New high-throughput methods are able to increasingly produce data very rampantly that can reflect the whole genome of an organism or a species. Due to this technological progress, it is not surprising that the volume of collected biological data is significantly increasing. Therefore, bioinformatics tools are necessary which cope with representation, modeling, analysis and visualization.

In this abstract, we introduce a new bioinformatics workflow which helps researchers to work and manage multidimensional data from different Omics- level in the form of biological networks. We aim to extend knowledge on the structural analysis of biological networks and hope to get fundamental insights into the underlying biological systems of integrated life science data.

Methods and Algorithms: The analysis is performed on experimental data and biological networks provided by the software application VANESA (http://vanesa.sourceforge.net/). VANESA is built upon the DAWIS-M.D. data warehouse [1] which is based on the BioDWH [2]. The analysis is performed on following integrated life sciences databases: BRENDA, HPRD, and ENZYMES. Since analysis methods on large networks are highly computationally intensive, we use parallel computing techniques on our cluster. Our computer cluster uses the aforementioned databases and architectures as a base for the now presented techniques. We have developed and implemented new methods and algorithms, which are parallelized, by using MPI. Algorithms such as shortest-path betweenness centrality measurement, node ranking, and topological structure analysis are the backbone of our architecture.

Results: We are able to generate and analyze networks that contain more than 28 million entries and 62 millions of connections.

Conclusion: Our first results are very promising and will truly help to analyze and understand the topological structures of large networks.

Availability: Available on request from the author.

References:

1. K. Hippe et al. (2010) DAWIS-MD-A Data Warehouse System for Metabolic Data. *GI Jahrestagung* **2**:720–725.
2. B. Kormeier et al. (2011) Data Warehouses in Bioinformatics: Integration of Molecular Biological Data. *it - Information Technology (IT),* **53**:241–249.

A NEW APPROACH TO IDENTIFY THE rSNPs IN THE HUMAN GENOME BASED ON CHIP-seq DATA

Bryzgalov L.O.*, Antontseva E.V., Matveeva M.Yu., Kashina E.V., Shilov A.G., Bondar N.P., Merkulova T.I.
Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
e-mail: leon_l@bionet.nsc.ru
** Corresponding author*

Key words: *rSNP, ChIP-seq, regulatory region, human genome*

Motivation and Aim: The functional interpretation of non-coding disease-associated single nucleotide polymorphisms (SNPs) identified by genome-wide association studies (GWAS) is challenging. Many of these SNPs are likely to be regulatory SNPs (rSNPs): variations which affect the ability of a transcription factor (TF) to bind to DNA. However, experimental procedures for identifying rSNPs are expensive and labour intensive. In silico methods are required for rSNP prediction usually based on weight matrix (PWM), and can determined which SNPs are likely rSNPs but its not very well. We report a new approach for rSNP prediction based on ENCODE ChIP-seq data analysis.

Methods and Algorithms: ChIP-seq data sets were downloaded from ENCODE project site UCSC. Potential rSNPs were searched in the ChIP-seq peaks overlappings. SNP localized in more than 8 ChIP-seq peaks were introduced into EMSA experiments. Protein nuclear extracts from the different human cell lines (HeLa S3, HTC, K562 and HepG2) were prepared for the EMSA.

Results: Genomic data on the locations of SNPs in the human genome were downloaded from database dbSNP NCBI http://www.ncbi.nlm.nih.gov/snp/ (Human Genome Build 37). Sample Sclinic consisting of 4046 clinical submitted SNPs was selected from dbSNP NCBI by limits criteria "organism: Homo sapiens" and "annotation: Clinical/LSDB Submissions" with following exception of SNPs having references on the proteins. Sample Somim composed of 3160 SNPs was also selected from database dbSNP NCBI in criteria "organism: Homo sapiens" and "annotation: OMIM" eliminating every SNPs having references on the proteins.This samples of SNPs were analized. 799 and 438 SNPs are localized in at least one ChIP-seq peak, accordingly. 110 and 64 SNPs fell at more then 8 ChIP-seq peak. For experimental verification 41 SNPs were selected. It was established that in the EMSA 31 SNPs influence on the binding of transcription factors. Subsequent analysis of the different samples of SNPs from dbSNP showed that samples Sclinic and Somim are enriched by SNP localized in ChIP-seq peak compared to samples of random SNPs. The relative enrichment was rising when we increased number of ChIP-seq peak overlaps in the SNP location.

Conclusion: We developed a new method for rSNP prediction. Experimental verification showed high efficiency of this method. We suggest that it may be used for regulatory SNP prediction.

POPULATION STUDY OF THE VARIATION IN TRIPLET DISTRIBUTIONS OBSERVED ALONGSIDE A CHROMOSOME, FOR YEAST SPECIES

Bushmelev Eu.Yu.
Siberian federal university, Krasnoyarsk, Russia
e-mail: eugenijbushmelev@gmail.com

Key words: *order, periodicity, correlation, function, taxonomy*

Motivation and Aim: The aim of the study is to identify, describe and visualize both the intragenomec, and intergenomec differences in triplet distributions observed alongside a DNA sequence, for a family of yeast genomes.

Methods and Algorithms: The distribution of the triplets alongside a sequence was developed. It was defined as a distance to the nearest neighbour, where two triplets ω_1 and ω_2 are as far, as n_l nucleotides one each other so that there is no other word ω_2 embedded somewhere inside the string of the length n_l. The longest distance to detect the nearest neighbour was as long, as 10^5 nucleotides. The distribution function was developed for all 4096 couples of triplets, for each chromosome. Then standard techniques of statistical analysis have been implemented to figure out the correlations and interdependencies between the distributions observed in the different chromosomes of the same species, and different species.

Results: A number of chromosomes of yeast genomes have been studied. All chromosomes exhibit a strong and extremely unusual structures in the distribution of the triplets (to the nearest neighbour). Thus, an explicit and strong periodicity in CCC – GGG triplets has been found, with the period of 13. Some other couples exhibit more complex and long-range correlations (up to 250 nucleotides). There is significant correlation in the distribution patterns observed within a genome, and quite sounding divergence in the correlations, when compared the chromosomes from different (while closely related) species, or stains.

Conclusion: a new character is figured out to study the evolution processes of DNA molecules. It is evident, that the found behaviour of the triplet distribution yields some universal patterns, and some specific ones. Such specific patterns can be reliably related to taxonomy of a genome.

LATENT STATISTICAL ORGANIZATION OF CODING AND NONCODING REGIONS IN HUMAN GENOME

Chaley M.B.*[1], Kutyrkin V.A.[2]
[1] *Institute of Mathematical Problems of Biology RAS, Pushchino, Russia;*
[2] *Moscow State Technical University n.a. N.E. Bauman, Moscow, Russia*
e-mail: maramaria@yandex.ru
** Corresponding author*

Key words: *latent profile periodicity, spectral-statistical approach, recognition of DNA coding regions*

Motivation and Aim: Development of current techniques for genome sequencing has led to data accumulation gets sufficiently ahead a possibility of their experimental and theoretical analysis. As the result, a great volume of the sequences considered as potentially coding DNA regions and waiting verification of their reliability is stored in the databases. Earlier a regularity in structural organization of the sequences from DNA coding regions was noted, which has occasionally been revealed by Fourier, correlation functions spectra and the like. Such incidentals did not allow elaborating reliable criteria for characterization of the coding regions. The present work is aimed to reliable revelation of specific latent regularity in the coding regions of human genome to discriminate them from the noncoding ones.

Methods and Algorithms: Methods and algorithms of spectral-statistical approach [1] are used in the work. This approach has been elaborated for recognizing regularity in DNA sequences along with a new type of latent periodicity – profile periodicity (profility).

Results: In the result of quantitative analysis of statistical properties of human DNA sequences a classification of regular structural characteristics has been done for coding and noncoding regions. This classification displayed cardinal differences between the coding regions and the introns in the sequences of human genome.

Conclusion: On the base of the classification, reliable statistical criterion has been proposed for selecting the coding regions. The criterion has been additionally tested on hypothetical proteins. In the majority of human genome coding regions (75% CDS) the latent profile periodicity was revealed and two-level organization was revealed for 13% CDS. Comparative analysis of Fourier methods and spectral-statistical approach techniques revealing the latent regularity in DNA sequences showed the higher sensitivity of the latter.

References:

1. M. Chaley, V. Kutyrkin. (2011) Profile-statistical periodicity of DNA coding regions, *DNA Res*., **18:** 353-362.

COMPREHENSIVE ANALYSIS OF UNIDENTIFIED LC-MS FEATURES FOR INVESTIGATING PROTEINS DIVERSITY IN HIGH-THROUGHPUT PROTEOMICS EXPERIMENTS

Chernobrovkin A.L.*, Zgoda V.G., Lisitsa A.V., Archakov A.I.
Institute of Biomedical Chemistry RAMS, Moscow, Russia
e-mail: chernobrovkin@gmail.com
** Corresponding author*

Key words: *single amino-acid polymorphisms; lc-ms; proteins identification*

Motivation and Aim: More than 65 thousands nsSNP are known to exist in human genome, and more than 20% of them associated with different diseases. However, the vast majority of annotated nsSNP have not been observed at protein level yet. Investigation of diseases-related nsSNP at protein level can shed light on the molecular nature of diseases and provide additional information for molecular biomarkers discovering.

Methods and Algorithms: According to recent estimation only a small proteomes can be analyzed properly using high-accuracy LC-MS without using MS/MS for peptide identification [1]. Within the human proteome only 20% peptides can be properly identified using only accurate parent mass and retention time data. Here we propose the new strategy for unidentified LC-MS features analysis, which allows significantly increase the sequence coverage of proteins, identified using MS/MS data and reveal protein variants caused by translation of non-synonymous nucleotide polymorphisms. The method uses accurate m/z and retention time data analysis for assigning theoretical peptides of identified using MS/MS proteins to the unidentified LC-MS features. As an additional resource for removing the ambiguity in features annotating we use quantitative data of protein abundance changes during cells differentiation.

Results: There were 1370 proteins identified in HL60 cells using LC-MS/MS (LTQ Orbitrap Velos, Thermo Scientific) analysis of triptically digested cell lysates. Quantitative analysis was performed using Progenesis-LC-MS software and allows us to reveal 300 proteins that have changed their abundance more than 3 times during cells differentiation process. LC-MS chromatograms were reanalyzed to select those features that could be matched to the triptic peptides of selected proteins and their variants. Such procedure allows two to three fold increase in the sequence coverage of selected proteins. Additionally we observed 38 features that match 17 SAP-specific proteotypic peptides of identified proteins.

Conclusion: Proposed approach makes it possible to decrease number of unsigned features in LC-MS based proteomics experiments. Assigning of additional features to previously identified proteins allows increasing protein sequence coverage and revealing variant-specific proteotypic peptides.

References

1. Bochet P. et al. (2010) Fragmentation-free LC-MS can identify hundreds of proteins, Proteomics, 11(1): 22-32.

CONTRIBUTION OF GENOTYPE VARIATION TO WARFARIN PHARMACOKINETICS

Chernonosov A.A.*, Koval V.V., Koh N.V., Tsvetovskaya G.A., Lifshits G.I., Fedorova O.S.
Institute of Chemical Biology and Fundamental Medicine SB RAS, Novosibirsk, Russia
e-mail: sandy@niboch.nsc.ru
** Corresponding author*

Key words: *warfarin, pharmacokinetics, mass-spectrometry*

Motivation and Aim. Warfarin is the one of the is one of the most widely prescribed oral anticoagulants worldwide; it is used to prevent and treat venous or arterial thrombi and emboli associated with atrial fibrillation or cardiac valve replacement. For the effective warfarin therapy patients have to take the lowest warfarin dose required to maintain the target international normalized ratio (INR). However, the individual doses required to each patients could be very different. It is known that cytochrome P450 (CYP), mainly CYP2C9, activity is an essential source of dosage variability. In addition, vitamin K epoxide reductase complex subunit 1 gene (VKORC1) plays a major role in the optimization of warfarin dose. The aim of this study was to assess the pharmacokinetics of warfarin in relation to cytochrome P450 (CYP2C9) and VKORC1 genotypes in residents of the West Siberian region.

Methods. The method was applied to determine the plasma concentrations of warfarin from a clinical trial in which 11 healthy volunteers received a single 5 mg oral dose of warfarin. All volunteers gave their signed informed consent to participate in the study. They have been genotyped for CYP2C9*2, CYP2C9*3 alleles and 1173C>T VKORC1 polymorphism. Blood samples were collected before and 1, 2, 3, 8, 24, 48 and 72 h post-dosing. Samples were centrifuged and plasma was separated and stored at −20 °C until analyzed. The concentrations of warfarin in the all plasma samples were analyzed by MS method using mass-spectrometer Agilent 6410 QQQ (Agilent Technologies, USA). Every samples were analyzes 3 times.

Results. All volunteers have *1/*1 genotype CYP2C9. Two of them have CC, five have TC and four have TT genotype VKORC1. For the TT and CT genotypes it was noted that $T_{1/2}$ of the warfarin is about 25% longer in compare with the CC VKORC1 genotype. Area under the curve (AUC) of warfarin concentration increased in the order CC, CT, TT genotypes VKORC1. Ratio of AUC between CC and CT genotypes was 1.65 and between CC and TT genotypes - 2. We have also observed that the clearance was increased in the order TT, CT, CC. It should be noted that this effect is also additive, and that heterozygotes respond to an intermediate warfarin dose, and homozygous carriers of the T allele respond to the lowest dose of warfarin. The results indicated that pharmacodynamic response to warfarin is highly variable between subjects. The association between warfarin concentration in plasma and 1173C>T VKORC1 genotypes suggests that the VKORC1 genotype significantly affects the personal dosage of warfarin.

Acknowledgements. Supports by grants from Ministry of Education and Sciences (no. 16.512.11.2073, NS-64.2012.4).

ANALYSIS OF TRANSCRIPTIONAL AND POSTTRANSCRIPTIONAL REGULATION OF AUXIN CARRIER *AtPIN1*

Chernova V.V. [1, 2], Ermakov A.A.[1], Doroshkov A.V. [1], Omelyanchuk N.A.[1], Mironova V.V. *[1]
[1] *Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia*
[2] *Novosibirsk State University, Novosibirsk, Russia*
** Corresponding author: kviki@bionet.nsc.ru*

Motivation and Aim: Auxin is the plant hormone which affects cell division, growth and differentiation. The family of PIN-FORMED (PIN) genes encode transmembrane proteins that transport auxin out of cells. PIN proteins are located asymmetrically in the plasma membrane of cells, thereby forming the tissue gradients and auxin concentration maxima affecting growth. It was shown earlier [1] that depending on the dose auxin has different effects on PIN1 expression in a cell. Our aim was to study the mechanisms underlying auxin regulation of PIN1 expression experimentally and by method for reconstruction of associative networks and interactions from published data.

Methods: PIN1::PIN1-GFP *Arabidopsis thaliana* seedlings were grown in a 16 hours light/8 hours dark cycle at 18-25°C on 1/2MS with sucrose. Short-time exogenous indole-3-acetic acid (IAA) application was performed by incubation of 4-5-day-old seedlings in liquid 1/2MS supplemented with different IAA concentrations (0,01 μM/l, 0,1 μM/l, 1 μM/l, 10 μM/l, 50 μM/l). Microscopical analysis of the plants was performed 24 hours after treatment. The experimental images were analyzed using ImageJ program. For reconstruction of interaction networks the program ANDCell [2] and databases IHop (www.ihop-net.org) and GeneMania (http://genemania.org) were used.

Results and conclusions: Experimental data on the expression of PIN1::PIN1-GFP allowed to estimate the following auxin-dependent changes in PIN1 protein expression: (i) increased PIN1 protein expression at low IAA concentrations and decreased one at high (> 1 μM / l); (ii) an increased PIN1 protein expression domain at low IAA concentrations and a decreased one at high (> 1 μM /l); (iii) treatments with low IAA concentrations (<0.1 μM /l) caused ectopic PIN1 protein expression in the epidermis and endodermis of root meristematic zone.

Treatment with high IAA concentrations (> 10 μM / L) resulted also in other alterations of PIN1 expression associated with root anatomical changes. Comparison of these data with [1] revealed differences in the expression of RNA and protein PIN1, including those which have not been previously described. We proposed that at high auxin concentrations in the root at least one of the following mechanisms may be switched on (1) PIN1 protein increased degradation, (2) inhibition of PIN1 translation; (3) suppression of PIN1 exocytosis and/or activation of endocytosis. We conducted analysis of published data on PIN1 expression using ANDCell [2], IHop and GeneMania software. As a result, the possible regulatory pathways have been found for suppression of PIN1 expression by high auxin concentrations, some of which will be further experimentally tested.

Acknowledgements: Microscopy was performed in the Shared Facility Center for Microscopic Analysis of Biological Objects SB RAS. The work is partially supported by the Dynasty Foundation grant for young biologists, RAS program A.II.6, Integration project SB RAS 80 and RFBR grants 10-01-00717-a, 11-04-01254-a.

References

1. Vieten et al. (2005) Development 132, 4521-4531.
2. Podkolodnaya et al., (2010) Herald Vogis. V. 14. N 1. P.106-115.

FUNCTIONAL ANALYSIS OF PUTATIVE TUMOR SUPPRESSOR GENES KCNRG AND KCTD7

Choi H.[1], Murthy S.B.K.[1], Baranova A.V.[1, 2]
[1] School of Systems Biology, George Mason University, Fairfax, VA, USA;
[2] Federal State Budgetary Institution "Research Centre for Medical Genetics" under the Russian Academy of Medical Sciences, Moscow, Russia
e-mail: abaranov@gmu.edu
** Corresponding author*

Key words: *tumor suppressors, analysis of correlations, functional genomics*

Motivation and Aim: KCNRG is a soluble protein with characteristics suggesting it forms hetero-tetramers with voltage-gated K+ channels (K_v) and inhibits their function. However, KCNRG related proteins do not bind voltage-gated K+ channels, but are associated with ubiquitin ligase cullin 3, suggesting that the function of KCNRG may be different from what was previously hypothesized. Cullin 3 ubiquitination is suspected to directly modify the activities of voltage-gated K+ channels. The KCTD7 gene is a paralog of the KCNRG gene that also binds to cullin 3.

Methods and Algorithms: The ONCOMINE database is an online collection of microarrays that profile various types of human cancer samples. It contains collections of microarrays analyzed in individual studies. Hundreds of tumor samples are described as a single, co-processed multi-array study to allow analyses of co-expression patterns. Separate analyses of ten different ONCOMINE datasets for co-expression patterns for the top 100 genes co-correlating with KCNRG and KCTD7 were performed, focusing on the brain and tissues of the central nervous system.

Results: The meta-analysis of the datasets with genes with frequency of 3 or above yielded 95 gene hits for KCNRG and 37 for KCTD7. These were further assessed for their full gene names and function ontologies. The Entrez IDs for all the overlapping genes were found using the Stanford University's SOURCE tool. This data was used as an input file for advanced analysis using Metacore, an integrated software suite for pathways and network analysis of OMICs data. Metacore is based on the curated database of human protein-protein and protein-DNA interactions, transcription factors, signaling and metabolic pathways, and diseases and toxicity.

The "Analyze Single Experiment" workflow in Metacore was employed for the meta-analysis of the data using 650 Canonical Pathways maps. This analysis showed that the top scoring map (map with the lowest p value) based on the enrichment distribution for genes co-expressed with KCNRG is "Transport RNA regulation pathway", and the top scoring map for KCTD7 was revealed to be "Cadherin-mediated cell adhesion". Further research is currently in progress.

Conclusion: Analysis of correlations may be applied to derive possible functions for novel human proteins.

A BACTERIAL MEMBRANE "ACHILLES HEEL": HIGH-PERFORMANCE COMPUTER SIMULATION OF PEPTIDOGLYCAN CARRIER LIPID-II IN THE CHARGED LIPID BILAYER

Chugunov A.O.*[1], Pyrkova D.V. [1], Nolde D.E.[1], Pentkovsky V.M.[2], Efremov R.G.[1]
[1] *Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry RAS, Moscow, Russia;*
[2] *Moscow Institute of Physics and Technology (State University), Dolgoprudny, Moscow region, Russia*
** Corresponding author. E-mail: batch2k@yandex.ru*

Motivation and aim. Unique conservative feature of the bacterial plasmatic membrane is cell-wall precursor called "lipid II". Being essential for growth of bacteria, lipid II is targeted by a wide class of bacteriocins called lantibiotics — ribosomally synthesized and post-translationally modified peptides with strong antimicrobial potency. Lantibiotics trap solvent-exposed part of lipid II molecule ("head") and inhibit cell-wall synthesis and (optionally) destroy the bacterial membrane. This recognition relies on both unusual conformational dynamics of lantibiotics and conservative structure of lipid II that is unlikely to change and develop resistance. In fact, lipid II is an "Achilles heel" of the bacterial membrane. Here, we study spatial structure of lipid II in model lipid bilayers and its effect on the membrane microenvironment using long-term molecular dynamics (MD) simulations.

Methods and Algorithms. The system under study consideration (≈5.5Ч10^4 atoms) contained a single lipid II molecule immersed into the hydrated "bacterial" anionic membrane (75% POPG / 25% POPE; a total of 288 phospholipids (PLs)). Several long-term (>420 ns) MD trajectories were collected with GROMACS software. Detailed atomic-scale mapping of the membrane surface (landscape and hydrophobic/hydrophilic properties) and analysis of lipid/lipid and lipid/solvent interactions were employed to estimate in a quantitative manner the effect of lipid II on the membrane surface.

Results. Our analysis revealed that lipid II molecule substantially deforms lipid bilayer (as compared to the membrane without lipid II): its head elevates out of the membrane surface with a prominent "ditch" around. Hydrophobicity mapping demonstrates that the lipid II head (consisting of pyrophosphate, two sugars and five modified amino acid residues) is very hydrophilic, while interactions with PLs make the later ones expose their acyl tails and form an atoll-like "hydrophobic ring" around the lipid II head.

Conclusion. Besides anchoring of lantibiotics directly on the solvent-exposed "head" of lipid II, the "defect" created by lipid II in the bilayer membrane may be another thing that makes bacterial membrane vulnerable to this class of antibiotics. Further study of this effect may contribute to design of novel potent analogues of lantibiotics.

AGROBACTERIUM-MEDIATED EVOLUTION?

Chumakov M.I.*, Mazilov S.I.
Institute of Biochemistry and Physiology Plants and Microorganisms, Russian Academy of Sciences, Saratov, Russia
e-mail: chumakov@ibppm.sgu.ru
** Corresponding author:*

Key words: *Agrobacterium, DNA transfer, evolution*

Motivation and Aim. Soil bacteria of the genus *Agrobacterium* are a natural vector for the ssT-DNA transfer into the eukaryotic cell genomes. For the first time a sequences similar with agrobacterial T-DNA was found for nontransformed *Nicotiana glauca* [1] and *N.tabacum* [2]. The data strongly suggest that an *Agrobacterium* infection resulted in genetic transformation early in the evolution of the genus *Nicotiana*. The aim of this study was to find new examples of the presence of *A.tumefaciens* T-DNA fragments in the other plants, animal and insect genomes.

Methods and Algorithms. T-DNA full sequences originated from *A. tumefaciens* Ti plasmids in GenBank using tblastn and blastn programs were evaluated.

Results. A set nucleotide sequences similar to the *mas1* (mannopinesyntetase), *tms2* (indolacetamidehydrolase), originated from *A. tumefaciens* T-DNA was found at 23 plant genomes: *Arabidopsis, Arachis, Artemisia, Brassica, Cajanus, Capsicum, Daucus, Jatropha, Lotus,.Medicago, Nicotiana, Oryza, Phyllostachys, Physcomitrella, Picea, Populus, Ricinus, Selaginella, Solanum, Sorghum, Triticum, Vitis* and *Zea*. We found *A.tumefaciens* T-DNA nucleotide sequences *(mas1, tms1, tms2* genes) within *Strongylocentrotus purpuratus* (sea urchins), *Caenorhabditis* (a free-living nematode) and insects genomes (13 species belonging to *Drosophila* genus*, Pediculus humanus, Bombyx mori, Spodoptera littoralis, Helicoverpa armigera, Anopheles gambiae*). The nucleotide sequences similar to the protein C (GenBank ID NP862658.1), coded by *rolB A. tumefaciens*, were found within plant genomes: *Catharantus*, *Linaria*, *Nicotiana, Plumbago*, *Withania.* The *hyuA, hyuB*, coding α-, β-subunits (acetone carboxylase) and *arc* genes located just after the right border T-DNA *A.tumefaciens*, have high homology in a number of plant genomes.

Conclusion: We conclude that agrobacterial T-DNA-born genes spreading into the plant, marine animal and insect genomes are possible involved in the evolution process.

Availability: tblastn, blastn programs located at {{http://blast.ncbi.nlm.nih.gov/Blast.cgi }}.

References

1. F.White, D.Garfinkel, G.Huffman, M.Gordon, E.Nester. (1983) Sequence homologous to *Agrobacterium rhizogenes* T-DNA in the genome of uninfected plants, *Nature*, **301**: 348-350.
2. A.D.Meyer, T.Ichikawa, F.Meins. (1995) Horizontal gene transfer: regulated expression of a tobacco homologue of the *Agrobacterium rhizogenes rolC* gene, *Mol. Gen. Genet.* **249**: 265-273.

SEQUENCE AND ANNOTATION OF THE CHROMOSOME OF PROBIOTIC STRAIN *LACTOBACILLUS RHAMNOSUS* 24

Danilenko V.N.*, Poluektova E.U., Klimina K.M., Kjasova D.H., Chervinetz J.V., Malko D.B., Makeev V.J., Gusev F.E., Tyajelova T.V., Reshetov D.A., Rogaev E.I.
N.I. Vavilov Institute of General Genetics RAS, Moscow, Russia
e-mail: valerid@vigg.ru
** Corresponding author*

Key words: *chromosome sequencing, Lactobacillus rhamnosus, toxin-antitoxin*

Motivation and aim: The goal of this work is to determine complete nucleotide sequence of the genome of probiotic bacterium *L. rhamnosus* 24 and to identify essential genes underlying probiotic properties of this bacterium. The genetic modules of our particular interest are toxin-antitoxin systems[1] which can be potential regulators of bacterial probiotic properties. These systems have not yet been studied in lactobacilli.

Methods and Algorithms: The strain *L. rhamnosus* 24 was isolated from the oral cavity of a healthy child (collaboration with microbiology and immunology chair of Tver Medical Academy (Russia)). The biochemical, probiotic and biotechnology properties of the strain have been characterized. To elucidate further the potential for probiotic activity of the strain we performed the complete genome sequencing of the strain. We have prepared the libraries of fragmented genomic DNA and proceeded with sequencing according to Illumina's HiSeq2000 manufacturer's paired-end protocol with some modifcations. Series of contigs were assembled into complete genome sequence with SOAP denovo. Glimmer program was used for *ab initio* gene prediction. Functional genome annotation was transferred from orthogous genes which were predicted using UniProtKB database.

The cloning of toxin gene was performed into the expression vector pET32a and the gene construct was transformed in *E.coli* D3 strain cells.

Results. The genome of *L. rhamnosus* strain 24 consists of one circular chromosome 2 647 976 bp with ~46,0% GC context. Apparently, the strain has no plasmid DNA. The genome of strain 24 was compared with nine reference genomes of *L. rhamnosus* strains from the GenBank (http://www.ncbi.nlm.nih.gov/genbank/). In addition to annotated *L. rhamnosus* genes, the genes unique for the strain 24 have been identified. In the genome of *L. rhamnosus* 24 seventeen genes for sugar utilization, two genes related to the process of adhesion, two probacteriocin genes, and one serine threonine protein kinase gene were found. Remarkably, two toxin-antitoxin systems (*PemK-A1*$_{Lrh}$ и *YefM-YoeB*$_{Lrh}$*)* and one toxin gene (*relE*$_{Lrh}$) were detected. Cloning and expression of YefM toxin in *E.coli* cells provided the evidence for functional activity of this protein, i.e., the capacity to kill bacterial cells.

Conclusion: L. rhamnosus strain 24 is the first lactobacilli probiotic strain from Russian resources/collections with the known genome nucleotide sequence. The data obtained in this study will be explored to evaluate the probiotic potential of this strain in comparison to other strains of geographically different origin.

Availability: At the moment the nucleotide sequence of *L. rhamnosus* strain 24 is available on request from the authors.

References:

1. A. A. Prozorov, V. N. Danilenko. (2010) Toxin–Antitoxin Systems in Bacteria: Apoptotic Tools or Metabolic Regulators? *Mikrobiologiya*, **79** (2):147-159.

E3 LIGASE AND THE P53 FAMILY PROTEINS INTERACTION MODELING

Davidovich P.*, Tribulovich V., Rozen T., Barlev N., Garabadzhiu A., Melino G.
Saint-Petersburg State Institute of Technology (Technical University), Saint-Petersburg, Russia
e-mail: davidovich_p@mail.ru
** Corresponding author*

Key words: *new target*, *p53 family, E3 ligase, protein folding, macromolecular docking*

Motivation and Aim: The p53 family proteins play crucial role in cell signaling, DNA repair and apoptosis. From the therapeutic point of view it's very important to understand the mechanisms how the p53 family proteins regulate activation/repression of gene expression processes. These pathways are appropriately investigated concerning the p53 protein, but there is no yet a clear understanding of how the activity of p73 is inhibited in human cells. When the first interaction of p53 and E3 ligase-MDM2 takes place, p53 becomes ubiquitinated, the second interaction leads to the formation of a diubiquitinated form. This process goes until the polyubiquitin chain is formed. Polyubiquitin chain is the "signal" for the proteasomal protein degradation. The aim of this work was to analyze the mechanistic properties of the MDM2-dependent deactivation/degradation of p73 by analogy with p53.

Methods and Algorithms: p53 and p73 transactivation domains were built by ab initio structure prediction modeling with QUARK server service [1] and GROMACS software package [2]. The interaction ability and binding energies of p53, p73 and MDM2 protein complexes were investigated by the macromolecular (protein-protein) docking method [3].

Results: The protein-protein docking with existing experimental models and predicted 3D-structures showed that the binding energy of p73 to MDM2 is about 100 kJ higher than the interaction between p53 and MDM2.

Conclusion: The differences in binding energy of MDM2-p53 and MDM2-p73 complexes make us strongly suggest that the mechanisms are likely to be different. We assume that the equilibrium *unbound form* ↔ *bound form* is essentially shifted to the bound form in case of the p73-MDM2 complex. The stability of the formed complex does not let the polyubiquitination process to be completed. So, the proteasomal degradation of p73 doesn't take place. Yet, MDM2 inactivates transcriptional function of the p73 protein.

Availability: none

Acknowledgements: The Grant of the Government of the Russian Federation for State Support of Scientific Research under the Guidance of Leading Scientists in the Russian Institutions of Higher Education № 11.G34.31.0069.

References:

1. D. Xu, Y. Zhang. (2012) Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. Proteins, (in press).
2. Hess, et al. (2008) GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. J. Chem. Theory Comput., **4**: 435-447.
3. D.W. Ritchie, V. Venkatraman. (2010) Ultra-Fast FFT Protein Docking On Graphics Processors, Bioinformatics, **26:** 2398-2405.

MODELING OF PLANT KINESIN-8 MOTOR DOMAIN AND RECONSTRUCTION OF ITS L2 AND L11 LOOPS

Demchuk O.M.*, Karpov P.A., Blume Ya.B.
Institute of Food Biotechnology and Genomics NAS of Ukraine, Kyiv
e-mail: demom79@gmail.com, demom@univ.kiev.ua
** Corresponding author*

Key words: *kinesin-8, modeling, bioinformatics*

Motivation and Aim: Among various classes of kinesines the kinesines-8 are minor class associated with microtubules [1]. The kinesines-8 play an important role in regulation of MT length [2]. In case of disruption of kinesin-8 mitotic functions leads to formation of abnormally long microtubules. Therefore it causes a disturbance during the divergence of homologous chromosomes [3]. The three-dimensional structure modeling of plant kinesin-8 motor domain is an important step in research of influence of tubulins posttranslational modifications on the interaction of mentioned proteins *in silico.*

Methods and Algorithms: Kinesin-8 motor domain three-dimensional models were constructed by I-TASSER (IT) and Swiss-Model Workspace (SMW) servers. Superimposition of kinesin structures and homology modeling of L11 loop using as a template a similar loop of chain A from crystal 2P4N were carried out using SwissPDBViewer 4.0.1. 3D-models were optimized with amber3 ff using conjugate gradient algorithm. Models were validated by MolProbity server.

Results and Discussion: Using Swiss-Model Workspace we generated kinesin-8 model from the primary sequence of *Arabidopsis thaliana* F25I16.11 by application chain B from crystal 3LRE (2.2 E) as matrix. I-TASSER generated also 5 models for the same sequence. MolProbity test of optimized models help to identify the best SMW-structure in comparison with IT-models. All constructed 3-D structures were superimposed with each other and with kinesines from X-Ray PDB-structures 3LRE (chain B) and 2P4N (chain A). It revealed some faults in the models generated *in silico*. The most problematic areas were L2 and L11 loops, which were also missed in X-Ray structure of human kinesin-8 (3LRE). However, L11 loop is available in the structure of kinesin in crystal 2P4N. We have constructed by homology a region S243-L267 from kinesin-8 *A. thaliana* (corresponds L11 loop) using as matrix the tertiary structure of S235-S259 region from 2P4N. In order to replace the coordinates generated by the server, the next step was to insert coordinates of constructed region S243-L267 into the SMW-structure. The area L36-V53 corresponding to the kinesin-8 L2 loop had a two β-strands. This area was replaced by a similar coiled region from one of IT-models. This region had the best results on «MolProbity» server.

Conclusion: The reconstructed model of kinesin-8 from *A. thaliana* has been optimized and demonstrated good score values on «MolProbity» server. These data suggest an acceptability of given model for subsequent bioinformatics studies

References:

1. A. Unsworth, H. Masuda, S. Dhut, T. Toda (2008) *Mol. Biol. Cell*, 19: 5104–5115.
2. G. Goshima, R.D. Vale (2006) *Nat. Cell Biol*., 8: 913–923.
3. M.M. Wargacki, J.C. Tay, E.G. Muller et al. (2010) *Cell Cycle*, 12: 2581–2588.

RNA-SEQ IDENTIFICATION AND ANALYSIS OF GENES CONTROLLING ABIOTIC STRESS RESPONSE IN BUCKWHEAT

Demidenko N.V.*[1, 2], Penin A.A.[1, 2, 3], Logacheva M.D.[2, 3, 4]
[1] *Department of Genetics, Biological faculty, M.V. Lomonosov Moscow State University, Moscow, Russia;*
[2] *Evolutionary Genomics Laboratory, Faculty of Bioengineering and Bioinformatics, M.V. Lomonosov Moscow State University, Moscow, Russia;*
[3] *A.A. Kharkevich Institute for Information Transmission Problems, Russian Academy of Science, Moscow, Russia;*
[4] *Department of Evolutionary Biochemistry, A.N. Belozersky Institute of Physico-Chemical Biology, M.V. Lomonosov Moscow State University, Moscow, Russia*
e-mail: demidenkonatalia@gmail.com
** Corresponding author*

Key words: *RNA-seq, gene expression, buckwheat, abiotic stress*

Motivation and Aim: Buckwheat (*Fagopyrum*) is a widely-cultivated pseudocereal belonging to eudicot family *Polygonaceae*. Due to its origin of radiation, North of India, buckwheat remains sensitive to low temperatures and draught. By now the information about genetic control of stress response systems in buckwheat is almost absent. Recent advances of high-throughput sequencing and gene expression analysis offer new opportunities for the analysis of this important trait.

Methods and Algorithms: Some essential abiotic stresses (heat, cold, rich light, wounding, darkness and over-illumination) were taken into analysis. Total RNA from cotyledons of stress-exposed and control plants was extracted and cDNA was sequenced using Illumina HiSeq2000 platform. After sequencing reads from all samples were assembled into general reference for further expression analysis. Differentially expressed genes identification was provided by CLC Genomics Workbench software [ver 5.0.1].

Results: A set of genes expressed as a response on stress exposure was identified. Putative orthologs of *Arabidopsis thaliana* genes were found and their functional similarity between *Arabidopsis* and buckwheat is analyzed.

Conclusion: Obtained data will promote the further analysis of abiotic stress response of buckwheat, functional analysis of differentially expressed genes and understanding of evolution of plant abiotic stress response systems.

PROTEOMIC OF MYCOPLASMAS: NANOFORMING *MYCOPLASMA GALLISEPTICUM*

Demina I.A.*, Serebryakova M.V., Ladygina V.G., Rogova M.A., Kondratov I.G., Renteeva A.N., Govorun V.M.
Research Institute for Physical_Chemical Medicine of Ministry of Public Health of Russian Federation, Moscow, Russia
e-mail: idemina@mail.ru
** Corresponding author*

Key words: *Mycoplasma gallisepticum*, *adaptation, starvation, reversion, proteomics, 2D electrophoresis, mass spectrometry, real-time PCR*

The goal of this work was the creation of a model for the long persistence of *Mycoplasma gallisepticum* in depleted medium and under low growth temperature followed by proteomic study of the model. Nanoforms and revertants for *M gallisepticum* were obtained. Proteomic maps were produced for different stages of the formation of nanoforms and revertants. It is shown that proteins responsible for essential cellular processes of glycolysis, translation elonga”

tion, and DnaK chaperone involved in the stabilization of newly synthesized proteins are crucial for the reversion of *M. gallisepticum* to a vegetative form. Based on the current, data it is assumed that changes in the metabolism of *M. gallisepticum* during nanoforming are not post”mortal, thus *M. gallisepticum* does not transform into an uncultivable form but remains in a reversible dormant state during prolonged unfavorable conditions.

EXTRACTION OF QUANTITATIVE CHARACTERISTICS DESCRIBING WHEAT LEAF PUBESCENCE WITH A NOVEL IMAGE PROCESSING TECHNIQUE

Doroshkov A.V. *, Genaev M.A., Pshenichnikova T.A., Afonnikov D.A.
Department of Systems Biology, Laboratory of evolutionary bioinformatics and theoretical genetics, Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
e-mail: ad@bionet.nsc.ru
** Corresponding author*

Key words: *common wheat, trichomes, leaf pubescence, high-throughput phenotyping, computer image analysis*

Motivation and Aim: Plant leaf pubescence (hairiness) plays an important biological role in adaptation to the environment and displays a wide phenotypic variation. Pubescence in wheat is due to unicellular unbranched epidermal outgrowths (trichomes). However, this trait has always been methodologically difficult to phenotype. An important step forward has been taken with the use of computer technologies.

Methods and Algorithms: Computer analysis of a photomicrograph of a transverse fold line of a leaf is proposed for quantitative evaluation of wheat leaf pubescence. We developed the LHDetect2 algorithm for the extraction of number of leaf hairiness characteristics, such as the number of trichomes in the image (it tells about trichome density); trichome length and its distribution (it tells about trichome size); average trichome length (it tells about average trichome size).

Results: In this part, we carried out a detailed study of leaf pubescence morphology in cultivar Hong-mang-mai and line 102/00^{i}. Previously, leaf pubescence density in this hybrid population had been analyzed by counting trichomes under magnification. It was then found that both cultivars have the allelic gene that controls the development of trichomes. However LHDetect2 that more detailed differences within the hybrid population were identified. We were able to identify the set of phenotypic classes among the second generation hybrids by the characteristics of leaf hairines. This allowed us to made an assumption about the number of genes and their interactions that can modify the expression of leaf pubescence trait.

Conclusion and Availability: The results demonstrate that the proposed method is rapid, adequately assesses leaf pubescence density, the length distribution of trichomes, and the data obtained using this method are significantly correlated with the density of trichomes on the leaf surface. Thus, the proposed method is efficient for high throughput analysis of leaf pubescence morphology in cereal genetic collections and mapping populations.

LHDetect2 software is available at http://wheatdb.com/lhdetect2.

APPLICATION OF REPAIR ENZYMES TO IMPROVE THE QUALITY OF THE DNA TEMPLATE IN PCR AMPLIFICATION OF DEGRADED DNA

Dovgerd A.P. *, Zharkov D.O.
Institute Chemical Biology and Fundamental Medicine, SB RAS, Novosibirsk, Russia
e-mail: antdovgerd@gmail.com
** Corresponding author*

Key words: *base excision repair, ancient DNA, forensic DNA, PCR, DNA damage*

Motivation and Aim: Although DNA is the main carrier of genetic information in living organisms, this molecule is inherently unstable. When an organism is alive, its repair systems resist DNA damage, but when these processes cease working after death, the accumulation of lesions in DNA becomes irreversible. Analysis of damaged DNA may be problematic. For example, the efficiency of PCR is sharply reduced if the template is subjected to oxidation or apurinization. This is particularly actual in studies of "ancient DNA" and DNA in the forensic practice. We are developing a system in which DNA repair enzymes are used to improve the quality of degraded DNA templates before PCR.

Methods and Algorithms: In most cases, lesions in postmortem DNA are located opposite an undamaged base in the complementary strand [1, 2] and, therefore, can be correctly repaired. We are creating a kit that includes several major DNA glycosylases, an AP endonuclease, a DNA polymerase and a DNA ligase, the combined effect of which leads to repairing much of the damage in DNA samples prior to their use as PCR templates.

Results: We have developed model systems of degraded high-molecular DNA with predominance of different types of lesions and have shown that the efficiency of PCR is sharply reduced if the template is subjected to oxidation or apurinization. We have reconstituted a repair system for a major oxidative DNA lesion, 8-oxoguanine, with *E. coli* 8-oxoguanine-DNA glycosylase Fpg, *E. coli* AP-endonuclease Nfo, human DNA polymerase β and bacteriophage T4 DNA ligase. We also include thermostable translesion DNA polymerase in the PCR to efficiently overcome the residual damage in the amplification process.

Conclusion: We have developing a system in which repair enzymes are used to improve the quality of degraded DNA templates before PCR. The primary use of our system may be found in the analysis of forensic samples, highly processed food and "ancient DNA".

Acknowledgements: This work was funded by the U.M.N.I.K. program (№06U/02-10).

References:

1. Hoss M., Jaruga P., Zastawny T.H., Dizdaroglu M., Paabo S. (1996) DNA damage and DNA sequence retrieval from ancient tissues. Nucleic Acids Res., 24, p. 1304–1307.
2. Mitchell D., Willerslev E., Hansen A. (2005) Damage and repair of ancient DNA. Mutat. Res., 571, p. 265–276.

EXPERIMENTAL EXAMINING OF PROGNOSES *IN SILICO* TBP BINDING TO TATA BOX WITH SNP ASSOCIATED WITH HUMAN DISEASES

Drachkova I.A., Ponomarenko P.M., Savinkova L.K.*, Ponomarenko M.P., Arshinova T.V., Kolchanov N.A.
Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
e-mail: savinkl@mail.ru
** Corresponding author*

Key words: *disease, polymorphism, TATA box, TBP, affinity; in vitro, in silico*

Motivation and Aim: It is known that completion of the integration project on a sequencing of the human genome brought huge results. Including at different people a large number of SNPs which phenotype consequences are unknown was revealed. But effective methods of a prediction of functionally significant SNPs genes regulatory regions are absent till now. The complex analysis of regulatory SNPs of the human genes will allow to receive real estimates of their influence on health of the person, sensitivity to drugs and environment conditions. Therefore the aim of the present work was to perform a standardized experimental examination of the change in TBP/TATA affinity as predicated by our equation (Ponomarenko et al., 2008) for disease-associated SNPs in human TATA boxes.

Methods and algorithms: Recombinant human TBP, was expressed in the *E. coli* strain BL21 (DE3) – plasmid pAR3038 -hTBP was kindly presented by prof. Puhg (Pennsylvania, USA). Experimental verification of prognoses изменения equilibrium dissociation constants of complexes TBP with TATA-containing oligonucleotides, corresponding to "normal" and "mutant" variants, the traditional approach including titration of fixed quantity TBP by increasing quantities TATA-containing oligos (length 26 bp), is used. Values of equilibrium dissociation constants TBP\TATA counted by means of the program "OriginPro 8".

Results. In this work we defined equilibrium dissociation constants of the human TBP/TATA complexes in most standardized experimental conditions (hTBP - full-size molecule identical on amino acids structure to a natural molecule of the human TBP) characterizing TBP affinity to TATA-boxes gene promoters of healthy people and sick alpha-, beta - and delta- thalassemia different weight, an amyotrophic lateral sclerosis, hemophilia B Leyden , neurosis, anemia, lung cancer, etc. The received quantitative values - ln [KD] and their changes, δ, at the contents of SNP in TATA-box ($\delta = -Ln[K_D,TATA\ Mut] - (-Ln[K_D,TATA])$, reliable correlate with in silico prognoses (factors of correlation - 0.779 and 0.682, respectively), made on the basis of the our equation of equilibrium binding TBP\TATA.

Conclusion: Thus, from the received results it is visible that little change of TBP/TATA affinity (1.5-3 folds) associates with the increased risk of emergence different diseases. Reliable correlation (r=0.663) between prognoses of affinity change and experimental analysis testifies to applicability of the deduced equation of linkage balance TBP/TATA for a prediction of potentially significant regulatory SNPs in human genome.

This work was supported by grant 10-04-00462 from Russian Foundation for Basic Research, grant 119 (Siberian Branch of the Russian Academy of Sciences), grant B.27 "Biological diversity" RAN.

SEMICONDUCTOR SEQUENCING FOR LIFE

Dyer M.
Life Technology, USA
e-mail: matt.dyer@lifetech.com

Ion Torrent has pioneered an entirely new approach to sequencing that enables a direct connection between chemical and digital information and leverage decades of semiconductor technology advances. The result is the first commercial sequencing technology that does not use light, and as a result delivers unprecedented speed, scalability, accuracy, and low cost. In just the first year the Ion Torrent Personal Genome Machine (TM) has become the fastest selling sequencing platform. The throughput scaled 100X, from 10Mb to 1Gb, in just the first year and will scale another 100X in the next year with the new Proton (TM) sequencer, which will enable the single day $1000 human genome. Automated data analysis is driven by Torrent Suite, an open-source software suite that provides a simple and intuitive interface to streamline data analysis and provide results in minutes to hours, not days. Built on top of Torrent Suite is a flexible SDK that allows users to expand the analysis capabilities through the development and utilization of plugins and APIs.

HETEROGENIC DATA MINING AND COMBINING

Efimov V.M.*[1, 2, 3], Kovaleva V.Yu.[2]
[1] *Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia*
[2] *Institute of Systematics and Ecology of Animals, SB RAS, Novosibirsk, Russia*
[3] *Tomsk State University, Tomsk, Russia*
e-mail: efimov@bionet.nsc.ru
** Corresponding author*

Key words: *multidimensional scaling, Mantel test, two block PLS-analysis*

Motivation and Aim: Suppose a set of objects has two heterogenic descriptions, for example, quantitative, rank or qualitative traits, text sequences, similarity/dissimilarity coefficients matrices, geometric morphometry data etc.

Our goal is to evaluate congruence of these descriptions to each other and to combine them for further processing.

Methods and Algorithms: A matrix of similarity/dissimilarity between every pair of objects calculate for every descriptions. Every matrix embeds as a set of points in a low-dimensional Euclidean space. Two sets of points rotate for achieving maximum congruence to each other (two block PLS-analysis) [1].

Results: This technology is applied to molecular-genetic (GenBank) and craniometric descriptions of small rodents interspecific variability in Western Siberia [2]. Correlation between descriptions is unexpectedly high (r = 0.766), and the configurations of two sets of points in the relevant Euclidean spaces are similar to one another. Combined distance matrix and species tree are calculated.

This tree coincides with the currently accepted taxonomy except for a one species which apparently should be in a separate genus.

For a given set of species the nucleotide positions and craniometric traits are found that uniquely discriminate among families.

Similar results are obtained by combining of molecular-genetic and morphometric descriptions of srew interspecific variability in Siberia and Far East.

Conclusion: Proposed technology is very promising and can be applied to solve very different bioinformatics problems.

Acknowledgements: The study was supported by RAS Presidium (Molecular and Cell Biology) Program 6.8, RAS Presidium (Biosphere Origin and Evolution) Program № 28, Scientific School NS-5278.2012.4, Integration projects of SB RAS № 63, 3.

References:

1. F.J.Rohlf, M.Corti (2000) Use of two-block partial least squares to study covariation in shape, Systematic Biology, **49:** 740-753.
2. V.Yu.Kovaleva et al. (2012) Analysis of congruence and a combining of molecular-genetic and morphological data in the zoological taxonomy, *Biological Bulletin*, **4.**

PROTEIN-PROTEIN AND PROTEIN-MEMBRANE RECOGNITION: A COMPUTATIONAL VIEW

Efremov R.G.[1], Polyansky A.A.[1], Chugunov A.O.[1], Nolde D.E.[1], Pentkovsky V.M.[2]
[1] *Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry RAS, Moscow, Russia;*
[2] *Moscow Institute of Physics and Technology (State University), Dolgoprudny, Moscow region, Russia*
e-mail: efremov@nmr.ru
** Corresponding author*

Motivation and aim. Cell membranes attract a growing attention as very perspective pharmacological targets. Rational design of new efficient and selective compounds modulating activity of biomembranes, requires atomic-scale information on their spatial structure and dynamics under different conditions. Because such details resist easy experimental characterization, important insight can be gained *via* computer simulations.

Methods and Algorithms. The work describes the results of computer simulations of structural/dynamic properties of membrane proteins and peptides with diverse fold, mode of membrane binding, and biological activities. Among the objects under study are: antimicrobial and cell-penetrating peptides, cardiotoxins from snake venom, transmembrane domains of receptor tyrosine kinases. The computational approach combines Monte Carlo simulations in implicit membranes, molecular dynamics in full-atom lipid bilayers, and molecular hydrophobicity potential analysis.

Results. Regardless different structure, dynamic behaviour, and mechanism of membrane permeation, in all cases the polypeptide-membrane recognition reveals a prominent "self-adapting" character. Namely, the membrane active agents employ a wide arsenal of structural/dynamic tools in order to insert into the lipid bilayer and to accomplish their function. Importantly, lipid bilayer of biological membranes plays an essential role in the recognition and binding events. In particular, the membrane surface reveals highly dynamic lateral heterogeneities (clusters), which differ in their packing and hydrophobic properties from the bulk lipids. Such a mosaic nature of membranes is tuned in a wide range by the chemical nature and relative content of lipids, presence of ions, etc.

Conclusion. Polepeptide-bilayer interactions represent a fine-tuned process, which requires the two active players – the polypeptide and the membrane. Interplay of the factors determining such a process assures efficient and robust binding of peptides and proteins to cell membranes. Understanding of such effects creates a basis for rational design of new physiologically active molecules and/or artificial membranes with predefined properties.

Acknowledgements: This work was supported by the Ministry of Education and Science of the Russian Federation, by the Russian Foundation for Basic Research, and by the RAS Programme "Basic fundamental research for nanotechnologies and nanomaterials". Access to computational facilities of the Joint Supercomputer Center RAS (Moscow) and Moscow Institute of Physics and Technology is gratefully acknowledged.

References:

1. Polyansky A.A. et al., (2009). J. Phys. Chem. B. 113, 1107-1119.
2. Pyrkova D.V. et al. (2011). Soft Matter 7, 2569-2579.
3. Konshina A.G. et al. (2011). PLoS ONE 6, e19064.

IRAP-PCR MARKERS AND MICRONUCLEI TEST IN THE CHARACTERIZATION OF GENETIC STRUCTURE OF THE KALMYK SHEEP AND TYPES OF THE EDILBAY SHEEP

Elkina M.A., Astafieva E.E., Glazko T.T.
Russian State Agrarian University–MTAA named after K.A.Timiryazev, Moscow, Russia
e-mail: mariyaelkina@yahoo.com

Keywords: *IRAP-PCR markers, micronuclei test, the Kalmyk sheep, the Edilbay sheep, the Suyunduk type, the Birlik type*

Motivation and Aim. The originality of the breed's gene pool is determined by the balance between the initial genetic diversity and populational and genetic responses to the factors of natural and artificial selection. In this regard, the selection of molecular genetic and cytogenetic methods that could identify the specificity of breed gene pool in order to control and optimize their use is of a particular interest. This is important especially for saving local breeds, such as the Edilbay and Kalmyk breeds of sheep.

Methods and Algorithms. The estimates of DNA polymorphism (IRAP-PCR markers) obtained using in polymerase cycle reaction (PCR) the terminal parts of retrotransposons: LTR SIRE-1 (GCA-GTT-ATG-CAA-GTG-GGA-TCA-GCA) , PawS 5 (AAC-GAG-GGG-TTC-GAG-GCC) and BARE-1 (CCA-ACT-AGA-GGC-TTG-CTA-GGG-AC) as the primers are used to study the genetic structure of populations, as well as the frequency of occurrence of erythrocytes with micronuclei (EMN in ‰) in the peripheral blood of two types of the Edilbay sheep (The Suyunduk and Birlik types) and the Kalmyk sheep.

Results. We obtained the following data. The Suyunduk type of the Edilbay sheep was the most homogeneous in comparison with the Birlik and Kalmyk sheep (P = 27%, PIC = 0,157; P = 29%, PIC = 0,221 and P = 47%, PIC = 0,199, respectively) according to the polymorphic information content (PIC) of the spectra of DNA amplification products (amplicons) obtained using PawS 5 primer in PCR. This type was also distinguished from the Birlic type and the Kalmyk sheep by its low frequency of EMN (0,2 ‰, 4,6 ± 0,3 ‰ and 4,3 ± 0,3 ‰, respectively). The Suynduk type differed from other groups of sheep by a high level of amplicon polymorphism in the case of use terminal fragments of retrotransposons LTR SIRE-1 and BARE-1 as primers in PCR. The relatively low frequency of EMN in peripheral blood of the Suyunduk type of the Edilbay sheep may be due to the fact that it was created in the nuclear testing area called the Azgirsky landfill, in contrast to other studied sheep groups reproducing in relatively more favorable environmental conditions. We could explain the relatively low frequency of cells with cytogenetic abnormalities in the Suyunduk type of sheep by chronic action of natural selection that promotes the reproduction of animals with increased resistance to environmental stress factors.

Conclusion and Availability. IRAP-PCR markers, using different terminal sites of retrotransposons as primers in PCR can reliably identify the specific features of gene pool of sheep breeds and intrabreed types. The characteristics of gene pool depend on the primers used in PCR. The Kalmyk sheep and the Birlik type of the Edilbay sheep did not differ from each other by micronuclei test. The relative reduction in the frequency of erythrocytes with micronuclei in a peripheral blood of the Suyunduk type suggests that a unique gene pool was created as a result of natural selection during chronic stressful environmental factors. That's why such populations have high potential for adaptation to adverse conditions of keeping and breeding.

BASE EXCISION REPAIR OF TRIPLET REPEAT SEQUENCES ASSOCIATED WITH NEURODEGENERATIVE DISORDERS

Endutkin A.V. *, Derevyanko A.G., Zharkov D.O.
Institute of Chemical Biology and Fundamental Medicine, SB RAS, Novosibirsk, Russia
e-mail: Aend@niboch.nsc.ru
** Corresponding author*

Key words: *8-oxoguanine, DNA glycosylases, base excision repair, trinucleotide repeats, dynamic mutations, neurodegenerative disease*

Motivation and Aim: Trinucleotide repeat expansion provides a molecular basis of several neurodegenerative diseases. One of the main reasons of triplet repeat expansion in somatic cells is base excision repair (BER), involving damaged base excision and DNA repair synthesis that may be accompanied by expansion of the repaired strand due to DNA looping. Expansion of CAG triplets characteristic of Huntington disease is initiated in the course of normal repair of the damaged base 8-oxoguanine (oxoG). Yet it is unclear how its repair might cause the expansion and whether the repair of other DNA lesions can lead to this expansion. Little is known about the efficiency of BER enzymes and their specificity when the DNA substrate contains trinucleotide repeats.

Methods and Algorithms: Using a number of CAG-substrates with the damaged triplet in different positions we have determined the rate constants of oxoG excision (k_2) and product release (k_3) by human oxoG-DNA glycosylase (OGG1) under conditions of single turnover kinetics and burst phase kinetics, respectively. Using fractionated cell extracts we have obtained the repair initiation rates for the same substrates. In the case of CAG-run substrates containing uracil in different position we have obtained the values of Michaelis constant and catalytic constants for the reaction of uracil excision by human uracil-DNA glycosylase (UNG). To determine the ability of the minimal BER enzyme set to repair oxoG and uracil in CAG-runs, we reconstituted the BER process *in vitro* with human DNA glycosylase OGG1 or UNG, AP endonuclease APEX1, and DNA polymerase β.

Results: We have analyzed kinetics of excision of ubiquitous oxidized bases, oxoG and uracil, by OGG1 and UNG from the substrates containing a CAG run. The values of k_2 rate constant for the removal of oxoG from triplets in the middle of the run were higher than for oxoG at the flanks of the run. The value of k_3 rate constant dropped starting from the third CAG-triplet in the run and remained stable until the 3'-terminal triplet in the run where it decreased even more. In nuclear extracts, the profile of oxoG removal rate along the run resembled the profile of k_2 constant, suggesting that the reaction rate in the extracts is limited by base excision. For uracil containing substrates, a strong dependence of both constants (k_{cat}, K_M) on the position of the damaged triplet was also observed.

Conclusion: The efficiency of initiation of repair of oxoG located in trinucleotide repeat runs depends on position of the damaged base in the run, namely, it is lower on its 5'- and 3'-borders. This dependence concerns both the rate constant of oxoG excision by pure DNA glycosylase OGG1 and the rate of cleavage of damaged DNA in cell extracts and efficacy of initial repair stages in the reconstituted BER system.

Acknowledgements: This work was supported by RFBR (10-04-91058-PICS_a) and Russian Ministry of Education and Science #14.740.11.1195.

GENOME SCANNING OF HORSE BREEDS BY USING OF ISSR-PCR MARKERS

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

Key words: *equus caballus, ISSR-PCR, multiloci scanning, genetic structure, genetic differentiation, Y chromosome marker*

Motivation and aim: Improving of breeding of local breeds is a very important task for agriculture because of their unique potential in biodiversity conservation and creation of gene pool reserves which is certainly necessary to create new breeds of agriculture animals. In current study we made an estimation of application of ISSR-PCR markers to reveal breed specific features of horse genetic structures: local breeds (Altaic horse breed, Karachay horse breed) and sport breeds (American trotter, Russian trotter, Orlov trotter).

Methods and Algorithms: A genomic scanning of 96 horses of Altai breed from two farms, 12 horses of Karachay breed and 48 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 summary, we observed 48 loci: 13 in $(AG)_9C$ spectra; 9 – $(GA)_9C$; 13 – $(GAG)_6C$ and 13 in $(CTC)_6C$ spectra. Each group had some specific loci. Two groups of Altai breed contained DNA fragments of 980, 900 and 450 bp length in $(AG)_9C$ spectra; 740 bp in $(GA)_9C$ spectra; 1180, 920 and 360 bp in $(GAG)_6C$; and 1500, 1430 and 1200 in $(CTC)_6C$. Trotter breeds had specific DNA fragments of 300, 250 and 180 bp length in $(GAG)_6C$ spectra. For Karachay breed only one unique fragment was observed: of 490 bp length in $(CTC)_6C$ spectra. The lowest PIC was found in Karachay breed (0.252), the highest – in altai breed (0.340). PIC in trotter breeds was very close to Karachay breed. Recently we showed that specie specific fragment from $(AG)_9C$ spectra may originate from recombination of evolutionary "old" and "young" mobile genetic elements [1]. We made primers to this fragment and found its presence only in males of investigated breeds. This fact allows supposing localization of this fragment in Y chromosome.

Conclusion: Genomic scanning made by using of ISSR-PCR markers allowed revealing of differentiation between genetic structures of breeds and interbreed horse crosses corresponding to their origin. Presence of amplicon of 416 bp length in $(AG)_9C$ spectra was observed only in males thus making it usable as Y chromosome marker.

References:

1. V.I. Glazko, A.V. Pheophilov, N.V. Barducov, T.T. Glazko. (2012) The species-specific ISSR-PCR markers and the ways of its appearance, *Izvestia TSHA*, **1**: 118 – 125.

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

Ermak T.[2], Timonov V.S. [2, 3], Akberdin I.R.*[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: akberdin@bionet.nsc.ru
** Corresponding author*

Key words: *kinetics data, parameters, E. coli, web database, biochemical reactions processes*

Motivation and Aim: One of the main challenge of the systems biology is a development of *in silico* cell. The cell is a theoretical platform (fundament) for performance of computational experiments and for verification of generated theoretical hypothesis on basis of the mathematical modeling. To reach this aim it demands development of robust mathematical models describing whole complex gene network of the prokaryote cell's functioning, for example. In turn, it requires accumulation and analysis of published information and data in web-warehouses on operation of the cell as an integrated system.

Previously, the Kinet database was developed in the ICG SB RAS [1]. The Kinet is a database of kinetics data and parameters of biochemical processes for *E. coli.* The database accumulates unique published data collected in the Institute over seven years. But the web access to the Kinet was unavailable for wide range of scientists.

Methods and Algorithms: Web application is developing on Java platform using Vaadin toolkit (http://vaadin.com).

Results: To solve the problem with availability to the Kinet we have developed the web application that has user friendly interface and easy access to unique data.

Conclusion: We have presented unique kinetics data of biochemical reactions (conditions of experiments, values of kinetics parameters, concentrations of key metabolites, cell enzymes and other) as the web-source. Researchers from all over the world can decrease time needed for development of mathematical modeling of biochemical processes, analyzing, selecting and other tasks related with kinetics data using the Kinet database. The Kinet will be helpful in carrying out of large-scale *in silico* studies. Also the database can be useful as information source.

Availability: Available upon request.

Acknowledgements: This study was also partially supported by the RFBR grant 11-04-01748-a, the Scientific school № 5278.2012.4 and Programs of the Presidium RAS "Molecular and cell biology" (6.8) as well as "Biological diversity" (30.29).

References:

1. Khlebodarova T.M *et al.*, Information sources for modeling *Escherichia coli* metabolic pathways // Proc. of the 5th BGRS International Conference, Vol. 2, pp. 19-24.

FOR LOOPS MODELING IN A GENOME

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

Key words: *genome structure, junk DNA*

Motivation and Aim: Design elements of the structure of the genome.

Methods and Algorithm: The coding of transcription factors (TFs) by genes which initiate the transcription of other genes can be considered as the possibility to pass control between the promoters of these genes. Unlike structural genes, which produce a large amount of a product, a control gene emits single TFs in the amount, necessary for a short-term activation of a controlled gene In contrast to the unconditioned transfer of control between genes, it is possible to construct a conditioned transfer of control. The gene, which takes the control, is activated only in the case of activation of two or more other genes. One of the mechanisms of such interaction can be described in the following way. TF or its precursor, emitting by the first control gene, is captured by a specific protein and can not activate the controlled gene. The second control gene produces an object, able to release the TF from the binding protein. In this case, the cooperation of two control genes results in the activation of the controlled gene. There can be some other mechanisms of the conditioned transfer of control, based on the existence of several binding sites for different TFs within the promoter of the controlled gene or on the presence of the corresponding epigenetic modifications of histones and chromatin in the promoter region of the controlled gene.

Results: The basic construction element of such "for loop" element is a calling module, or a "step". Loop step represents a gene, activating by the presence of two controlling effects and encoding two TFs. One of these TFs conditionally transfers the control to the next loop step; another one activates a certain subroutine, at the end of which one of the genes emits the releasing factor for the next loop step. Then the process is repeated. Each step activates the subroutine and after its completion transfers the control to the next step. Moreover, each step is able to run a unique subroutine, specific for this step. It can be simple and contain only several control and structural genes, or be complex, like programs, describing the life cycle and division of the cells. The step of the for loop of the genome is fundamental for multicellular organisms. It represents the gene, emitting three types of TFs and determining life cycle of the cell. The first and second TFs, blocked by specific proteins, are directed to the regions, in which the nucleuses of the first and second daughter cells will be formed, respectively. The third TF transfers control to the program of the maternal cell, using it as a subroutine for this step. The development, specialization, and division of the cell occur in accordance with this program. After the completion of the cell program, one of its genes emits objects, releasing transcription factors in daughter cells from the binding proteins. In the first daughter cell, the next loop step of the genome is initiated by the released first transcription factor. The second transcription factor, released in the second daughter cell, initiates another loop step of the genome. Then the process is repeated for each of the daughter cells. The step of the for loop of the genome can be introduced by two genes each of which encodes two TFs.

Conclusion: The most part of the genome of multicellular organisms consists of almost repeated elements - loop steps of the genome, and only one step demonstrates a short-time activity in each cell. Other steps remain inactive and look like junk DNA. This model was created by the author.

ENHANCER MODEL

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

Key words: *genome structure, enhancer*

Motivation and Aim: Design elements of the structure of the genome.

Methods and Algorithm: The coding of transcription factors (TFs) by genes which initiate the transcription of other genes can be considered as the possibility to pass control between the promoters of these genes. The functioning of a cell, including such stages as the cell growth, specialization, and subsequent cell division, is subjected to a unique program, representing an oriented graph, which fragments transfer the control from each other in a certain order and are able to call various subroutines, gene networks, and separate structural genes. Unlike structural genes, which produce a large amount of a product, the gene, which transfers the control, emits single TFs in the amount, required only for a short-time activation of the controlled gene. Under favorable conditions, the amount of the product, produced by a structural gene, is proportionate to the amount of a TFs, able to initiate its synthesis. In the case when control genes emit a small number of TFs, the process requires an "amplifier", i.e., a gene, which, being activated, is able to produce a large number of TFs for a controlled structural gene. The above-mentioned features are inherent to the enhancer.

Results: There can be several possible types of the enhancer structure.

1. The enhancer represents a gene, which coding region contains a large number of tandemly repeated copies of a TF precursor. During a post-transcriptional modification, this transcript is cleaved into a large number of fragments, which are later transformed into single TFs, able to activate a controlled structural gene.

2. The enhancer represents a gene, which coding region contains two different TFs; one of them is able to the repeated activation of its own promoter, and another is able to activate the promoter of a controlled structural gene. A single initiation of such enhancer makes it possible to obtain an unlimited number of TFs. If the coding region of such enhancer contains more than one copy of the second TF, then the enhancer productivity will be in direct proportion to the number of TF copies.

3. The enhancer represents a gene, which coding region contains a single TF, able to activate a controlled structural gene, or a certain number of its copies. At the same time, the splicing and the subsequent modifications of one of the introns of the controlled structural gene transform it into the TF, able to the repeated activation of the enhancer. As a result, a positive feedback is formed between the enhancer and the controlled structural gene; this feedback provides a continuous expression of the structural gene. The results of experiments, confirming this model, are discussed in [1].

Conclusion: In this model, the spatial approach of an enhancer to the promoter of a controlled gene, which was observed in many studies, should facilitate the TF exchange between these elements. The offered model is based on a hypothesis that TFs represent short non-coding RNAs, which, being a part of nucleoproteins, interact with the complementary sites of promoters and initiate the transcription of genes. This model was created by the author.

References:

1. Bing Ren."Transcription: Enhancers make non-coding RNA". Nature, Volume 465, Pages: 173–174,13 May 2010.

Gp39, A NOVEL PHAGE-ENCODED INHIBITOR OF BACTERIAL RNA POLYMERASE

Esyunina D.M.*[1, 2], Miropolskaya N.A.[1], Minakhin L.S.[3], Kulbachinskiy A.V.[1]
[1] *Institute of Molecular Genetics, Russian Academy of Sciences, Moscow, Russia;*
[2] *Molecular Biology Department, Biological Faculty, Moscow State University, Moscow, Russia;*
[3] *Waksman Institute of Microbiology, Piscataway, New Jersey 08854, USA*
e-mail: es_dar@inbox.ru
** Corresponding author*

Key words: *RNA polymerase, transcription initiation, open promoter complex, bacteriophage, transcription inhibitor*

Motivation and Aim: Bacterial RNA polymerase (RNAP) is an important target for antibacterial therapy. Bacteriophages often encode regulatory proteins that can bind and inhibit bacterial RNAP during phage infection. Such proteins and their derivatives can therefore be used for development of novel transcription inhibitors. Recently, we isolated a novel RNAP-binding protein, gp39, encoded by phage P23-45 that infects thermophilic bacterium *Thermus thermophilus*. This small 16 kDa protein was shown to inhibit transcription inihiation by RNAP. The purpose of this study was to characterize the mechanism(s) of transcription inhibition by gp39.

Methods and Algorithms: The effects of gp39 on different steps of transcription were characterized using *in vitro* transcription approaches with highly purified RNAP preparations from several bacteria. To localize the gp39-binding determinants, mutant gp39 and RNAP variants were obtained by site-directed mutagenesis followed by their expression, purification and functional analysis.

Results: Gp39 was found to affect different steps of transcription by bacterial RNAP. During initiation, gp39 specifically inhibits recognition of the -10/-35 class of bacterial promoters with high efficiency. During elongation, gp39 stimulates RNA synthesis and suppresses transcription termination. Gp39 was shown to target the flap domain of RNAP which is involved in both promoter recognition and RNA binding during transcription elongation. The observed effects of gp39 on initiation and elongation were shown to be independent of each other. Thus, gp39 is a bi-functional protein that exerts both inhibiting and activating effects on transcription. Analysis of the mechanism of inhibition of transcription initiation by gp39 revealed that it prevents the open complex formation, likely by disrupting recognition of the -35 element by RNAP. We propose that the interaction between gp39 and the RNAP flap domain can be used for development of fluorescence-based assays for selection of transcription inhibitors targeting the flap domain of RNAP.

Conclusion: Gp39 is a potent RNAP inhibitor that disrupts its interactions with promoters during transcription initiation. Gp39 can be used for design of small oligopeptide inhibitors of transcription and for development of highly sensitive assays for screening of novel antibacterial compounds targeting bacterial RNAP.

Acknowledgements: This work was supported by Federal Targeted Program "Scientific and scientific-pedagogical personnel of innovative Russia 2009-2013" (state contract 02.740.11.0771), by the Russian Foundation for Basic Research grant 10-04-00925 and by the grant of the President of Russian Federation MK-7156.2012.4.

DISCRETE AUTOMATON MODEL OF GENE NETWORK WITH VARIOUS FORMS OF REGULATORY ACTIVITY OF AGENTS (BASED ON *E. COLI*)

Evdokimov A.A.[1], Kochemazov S.E.*, Otpuschennikov I.V., Semenov A.A.
[1] *Sobolev Institute of Mathematics, SB RAS, Novosibirsk, Russia*
e-mail: evdok@math.nsc.ru
Institute for System Dynamics and Control Theory, SB RAS, Irkutsk, Russia
e-mail: veinamond@gmail.com
** Corresponding author*

Key words: *Gene networks, discrete models of gene networks, program translations, Boolean equations, SAT*

Motivation and Aim: First discrete models of gene networks were proposed by S. Kauffman in 1969 [1]. These models were based on Boolean networks. Paper [2] was one of the first works in which dynamical properties of Kauffman networks were studied using modern computational symbolic algorithms. In [3] there was proposed a class of discrete automaton models of gene networks with more complex structure than that of Kauffman networks. In [4] so called SAT-approach was used to study the properties of networks from [3]. In the present paper we propose a new class of models of gene networks similar to the one from [3] but with more complex form of regulatory activity of agents.

Methods and Algorithms: In proposed models we use more complex versions of threshold functions from [3] to describe the dynamical properties of gene network. To reduce corresponding problems to Boolean equations we use a special translator of algorithms. Obtained systems of Boolean equations are solved using modern SAT-solving algorithms.

Results: We tested the proposed model using a fragment of the regulatory circuit of *E.Coli* with 96 vertices. For the discrete automaton mapping obtained it was possible to find sets of fixed points in real time.

Conclusion: We believe that the model and methods proposed may have valuable applications to the research of gene networks with complex nature of regulation.

Availability: We used software complex TransAlg [5] to reduce problems considered to Boolean equations. For solving the latter we used SAT-solver MiniSAT 2.0 [6].

References:

1. S. A. Kauffman (1969), Metabolic stability and epigenesis in randomly constructed nets, *Journal of Theoretical Biology,* **22**, 437–467.
2. E. Dubrova et al. (2005) Kauffman Networks: Analysis and Applications, *Proc. of Int. Conf. on Computer-Aided Design (ICCAD'2005)*: 479-484.
3. A.A. Evdokimov et al. (2005) Nepodvizhnie tochki i cikly automatnih otobrazheniy modeliruyushih funkcionirovanie gennih setey, *Vestnik TSU*, **14**: 206-212.
4. A.A. Evdokimov et al. (2011) Application of symbolic computations to the study of discrete models of some gene networks (in russian), *Computational technologies*, **1**: 30-47.
5. I.V. Otpuschennikov, A.A. Semenov (2011) Technology for translating combinatorial problems into Boolean equations, *Prikladnaya Diskretnaya Matematika*, **1**: 96-116.
6. MiniSat [http://minisat.se/]

BRI-SHUR.COM – A SITE FOR BIOINFORMATICS COMPUTATIONS

Feranchuk S.I.*[1], Potapova U.V.[2], Potapov V.V.[2], Mukha D.V.[3]
[1] *Belarussian State University, Minsk, Belarus, e-mail: feranchuk@gmail.com;*
[2] *Limnological Institute SB RAS, Irkutsk, Russia;*
[3] *I nstitute of Bioorganic Chemistry NAS, Minsk, Belarus*
e-mail: feranchuk@gmail.com
** Corresponding author*

Key words: *web services, homology modeling, similarity screening, multiple alignment*

Motivation and Aim: The presented site is serving as a platform to provide a researcher with modern unique tools and pipelines in various subjects of computational biology.

Methods and Algorithms: Several original algorithms on similarity screening, multiple alignment, protein energy estimation, natural language processing and data mining, as well as publicly available software tools, are organized as web services and pipelines are established between them to achieve a maximal usability of the site. Some of the novel algorithms are originated from a Genebee group in Moscow State University.

Results: Scientific results include a good quality of pipeline on homology modeling, starting from a query sequence, in comparison with a well-known tool psi-blast, with a use of novel algorithms on similarity screening and protein energy estimation. Also several applied investigations were facilitated using the presented services. Search tools on Medline database of biomedical literature become popular in the Internet.

Conclusion: The algorithm processing natural language constructions allows to achieve a new level of actualization of scientific data on biomedical research accumulated in bibliographic databases. The site is designed in hope that it will be extensively used in bioinformatics community in Russia and everywhere.

Availability: Currently all the services are provided free of charge and without registration. The site is available at http://bri-shur.com

Acknowledgements: The authors would like to thank prof. A.V. Tuzikov from United Institute of Informatics Problems in Minsk and Dr. S.I. Belikov from Limnological Institute SB RAS for a support of their work.

MOSS PROTEOMICS AND PEPTIDOMICS. NEW INSIGHT IN THE OLD STORY. PEPTIDES IN THE STRESS ADAPTATION PROCESS

Fesenko I.A.*[1], Slizhikova D.K.[1], Seredina A.V.[1], Mageyka I.S.[1], Govorun V.M. [1, 2]
[1] *Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry RAS, Moscow, Russia;*
[2] *Institute of Physico-Chemical Medicine, Moscow, Russia*
e-mail: fesigor@gmail.com
** Corresponding author*

Key words: *Physcomitrella patens, proteome, peptidome, stress*

Motivation and Aim: The moss *Physcomitrella patens* is a new model system in plant science, that offers several advantages for studying of plant physiology, biochemistry and genetics. Phylogenetically, *P. patens* is situated in a key position between the green algae and the seed plants. The ancestors of mosses and seed plants separated shortly after the transition from water to land at least 500 million years ago. In addition, *P. patens* is the only land plant which has an exceptionally high rate of homologous recombination. We utilized the *P. patens* protoplast as a model to explore the mechanisms involved in stress adaptation in plant.

Methods and Algorithms: We used proteomic analysis to measure changes in the protein composition of freshly isolated protoplasts from the moss protonema. For this purpose, we compared results of 2D electrophoregrams of proteins from protoplasts and protonema using specific fluorescent dyes (DIGE) for identification of proteins specific to different living forms of *P. patens*. Besides, using a combination of high performance mass spectrometry with a bioinformatic analysis we described peptidome of the moss protoplast and green tissue.

Results: The DIGE demonstrate difference in the protein compositions of protoplasts and protonema. Thus, we have detected protein spots on two-dimensional electrophoregrams that were identified by mass spectrometry as Rubisco fragments in MW range from 10 to 20 kDa. The analyses of peptidome showed that the amount of peptides identified in protoplasts is almost six times greater than in the protonemata from which they are isolated and five times greater than in gametophores.

Conclusion: The isolation of moss protoplasts is accompanied by the degradation of proteins most of which are the proteins that belongs to the system of photosynthesis. Those processes of protein degradation lead to the generation of endogenous peptides, which is peculiar to stress responses of higher plants.

VALIDATION OF THE *PPP1R12B* AS A CANDIDATE GENE FOR CHILDHOOD ASTHMA SUSCEPTIBILITY

Freidin M.B.*[1], Polonikov A.V.[2]
[1] *Research Institute of Medical Genetics, Tomsk, Russia;*
[2] *Kursk State Medical University, Kursk, Russia*
e-mail: mfreidin@medgenetics.ru
** Corresponding author*

Motivation and Aim: Genome-wide association studies (GWAS) are a powerful tool for revealing positional candidate genes of complex diseases and traits. A number of such the studies have been performed in childhood and adult bronchial asthma (BA), which allowed disclosing dozens of novel candidate genes. However, given the high genetic heterogeneity of BA, a validation of the GWASs in independent populations is of critical importance. Recently, the *PPP1R12B* gene was revealed as a new candidate gene for childhood asthma in GWAS in Russians of West Siberia using Illumina 610QUAD chip (1). We now set out to validate this finding in Russian population of Kursk.

Methods and Algorithms: One hundred and fourteen patients with childhood BA (age-of-onset up to 16 years) and 279 healthy controls were genotyped using same Illumina chip. After quality assessment, total of 26 markers embracing the *PPP1R12B* gene region were analyzed using linear models approach to measure the associations between the disease and markers.

Results: Four markers (rs17438212, rs12734001, rs3767423, and rs3817222) were found to be significantly associated with childhood BA after correction for multiple testing. They represent two relatively distinct linkage disequilibrium blocks with two markers in each. The odds ratios for the associations varied between 1.78 and 2.04 (Padj = 0.029 – 0.03). The direction of association and the magnitude of the effect for the rs12734001 and rs3817222 markers were in accordance with initial finding in Russians of West Siberia, providing the validation of its verity.

Conclusion: Thus, the present data confirmed that the *PPP1R12B* gene is associated with childhood BA in Russians. Further studies are required to provide the functional basis for the association.

References:

1. M.B. Freidin et al. (2011) Genome-wide association study of allergic diseases in Russians of West Siberia, Molecular Biology, 45: 421-429.

EVOLUTION TRANSITION TO COMPLEX DYNAMIC MODES IN STRUCTURED BIOLOGICAL POPULATIONS

Frisman E.Ya.[1], Zhdanova O.L.*[2]
[1] *Complex Analysis of Regional Problems Institute FEB RAS, Birobidzhan, Russia;*
[2] *Institute of Automation and Control Processes FEB RAS, Vladivostok, Russia*
e-mail: axanka@iacp.dvo.ru
** Corresponding author*

Key words: *evolution, age structure, attractor, equilibrium, population size, stability*

Motivation and Aim: The deep understanding of activity results of intra-population self-organization mechanisms is necessary for further investigation the question: what occurs with biological population affected by changing factors of external environment? This work devoted to the analysis of connection between ontogenesis duration and mode of dynamic behavior of biological community in condition that only intra-population mechanisms of number growth regulation are considered. This work proceeds with research series of natural evolution in biological population with marked seasonality of life-cycle.

Methods and Algorithms: Mathematical and computer modeling of population dynamics.

Results: Investigation shows that attractors of fewer dimensions than maximal possible one are prevalent in the large part of acceptable parametric region of reproductive potential values. Although increasing of ontogenesis duration follows by growth of potential possibilities for intensification of systems dynamics chaotization but expected growth of chaotization does not occur and in average the dynamics of system with more complicated structure looks like less various then those of population with short ontogenesis. In biological aspect this fact means that population with long ontogenesis in average has more ordered dynamics and consequently it is more viable.

Conclusion: Growth of ontogenesis duration and complexity does not increase the power of attractors' chaotization. The most dynamic stability is presented by such factors as increasing of reproductive potential values region with stable dynamics in multi-age populations, the restriction of fluctuation scope of population groups' sizes, and scant diversity of attractors with large dimension. This result provides one possible explanation at the model level the fact that many natural biological population with age structure demonstrate clear stable or pseudo-cyclic dynamics [Freckleton, Watkinson, 2002; etc.], despite there are wide variety of dynamic regimes that theoretically possible for structured populations [Greenman, Benton, 2004].

*Acknowledgements: T*he present work was supported by the Russian Foundation for Basic Research (project 11-01-98512) and the Far Eastern Branch of Russian Academy of Sciences in the framework of Program 28 of Presidium of RAS (project FEB RAS 12-I-P28-02, 12-II-CO-06-019, 11-III-B-01M-001)

References:

1. R. P. Freckleton, A. R. Watkinson. (2002). Are weed population dynamics chaotic?, *Journal of Applied Ecology,* **39:** 699–707.
2. J. V. Greenman, T. G. Benton. (2004). Large amplification in stage-structured models: Arnold tongues revisited, *Journal of Mathematical Biology*, **48:** 647–671.

COMPLETE MITOCHONDRIAL AND CHLOROPLAST GENOMES OF DIATOM ALGA *SYNEDRA ACUS*

Galachyants Y.P.
Limnological Institute, Siberian Branch of the RAS, Irkutsk, Russia
e-mail: yuri.galachyants@lin.irk.ru

Key words: *diatoms, complete mitochondrial and chloroplast genomes, next-generation DNA sequencing, comparative genomic analysis, phylogenetic analysis*

Motivation and Aim: Only several complete mitochondrial and chloroplast genomes of marine diatoms have been sequenced to date. Here we present the complete mtDNA and cpDNA sequences of freshwater araphid pennate diatom alga *Synedra acus* subsp. *radians* (Kьtz.) Skabitsch from Lake Baikal. Also we present the results of comparative genomic and phylogenetic analyses for available diatom genome sequences.

Methods and Algorithms: To sequence mtDNA and cpDNA we analyzed the shot-gun genomic library prepared from *S. acus* total DNA with Roche/454 GS FLX Titanium instrument. The mitochondrial- and chloroplast-specific contigs were identified in the assembly according to their similarity with known organellar diatom sequences. Finishing of mtDNA was performed using the primer-walking approach. For chloroplast genome, the reference-guided assembly of the short Illumina reads was used to enhance the quality of cpDNA sequence. At final step, the overlapped ends of the chloroplast contigs were merged into a circular molecule.

Results and discussion: Mitochondrial genome of *S. acus* has length of 46,657 bp. It encodes 2 rRNAs, 24 tRNAs, and 33 proteins. The mtDNA of *S. acus* contains three group II introns which seem to be polyphyletic. The compact gene organization contrasts with the presence of a 4.9 kb-long intergenic repeat region. Comparison of the three sequenced mtDNAs showed that these three genomes carry similar gene pools, but the positions of some genes are rearranged.

Chloroplast genome of *S. acus* possesses a canonical quadripartite structure and maps as a circular molecule of 116 251 bp. It encodes 160 genes including tRNAs, rRNAs, and 128 protein genes. Comparative analysis of diatom cpDNA reveals 154 common genes and the absence of an overlapping between *atp*D and *atp*F gene coding sequences in *S. acus* genome. The transfer of *acpp* genes to a host nuclear genome is hypothesized to occur independently in several lineages of diatoms.

Acknowledgements: This work was financially supported by the Program of the Presidium of RAS ‘‘Molecular and Cell Biology’’ (project #22.3). We thank the distributed computing group of IDSTU SB RAS in help to perform the computationally-extensive bioinformatic analyses.

References:

1. N.V. Ravin *et al*. (2010) Complete sequence of the mitochondrial genome of a diatom alga *Synedra acus* and comparative analysis of diatom mitochondrial genomes, Curr. Genet., **56:** 215-223.
2. Y.P. Galachyants *et al*. (2012) Complete chloroplast genome sequence of freshwater araphid pennate diatom alga *Synedra acus* from Lake Baikal, Intl. J. Biol., **1:** 27-35.

IDENTIFICATION OF RARE VARIANTS AND POLYMORPHISMS OF THE *IL12RB1* GENE AND ANALYSIS OF THEIR ASSOCIATIONS WITH TUBERCULOSIS

Garaeva A.F.*, Rudko A.A., Bragina E.Yu., Babushkina N.P., Freidin M.B.
Research Institute of Medical Genetics, SB RAMS, Tomsk, Russia
e-mail: gaf_1986@mail.ru
** Corresponding author*

Key words: *atypical familial mycobacteriosis; IL12RB1, nucleic acids sequencing; rare variants, polymorphisms*

Motivation and Aim: Interleukin-12 acting through specific receptor encoded by the *IL12RB1* gene plays a key role in the immune response to *M. tuberculosis*. Several mutations in the *IL12RB1* gene lead to the development of rare syndrome of atypical familial mycobacteriosis (OMIM # 209950). This suggests that rare variants and polymorphisms of this gene predispose to tuberculosis.

The aim of this study is to find rare variants of the *IL12RB1* gene and evaluate their prevalence in TB patients and healthy Russians of Tomsk region.

Methods and Algorithms: The search for rare variants of the *IL12RB1* gene was carried out by Sanger's method using Applied Biosystems 3130xl Genetic Analyzer in 10 individuals suffered from aggressive forms of TB. The analysis of the results of sequencing was performed using BioEdit and Sequence Scanner v1.0 software. On the next stage of the study the identified nucleotide substitutions were genotyped in 310 patients and 250 healthy individuals using PCR-RFLP analysis. Association analysis was done using the http://ihg2.helmholtz-muenchen.de/ihg/snps.html on-line resource.

Results: Sequencing of intron-exon regions of the gene *IL12RB1* identified seven previously described polymorphisms, including four in exons (synonymous substitutions rs11086087, rs17852635; missense-mutations rs11575934 and rs401502), two in introns (rs12461312, rs17882555), and one in 3'-UTR (rs3746190). Synonymous substitutions were excluded from further analysis. Other variants were genotyped in patients and healthy individuals. The polymorphisms demonstrated no statistically significant differences between the studied groups.

Conclusion: No rare variants in the *IL12RB1* gene in patients with severe forms of tuberculosis were revealed by direct sequencing. The detected known polymorphisms in the *IL12RB1* gene are not associated with tuberculosis.

GENETIC BARCODE AS A PERSONAL IDENTIFIER OF EACH INDIVIDUAL

Garafutdinov R.R.*, Chubukova O.V., Sakhabutdinova A.R., Mashkov O.I., Shakirov I.G., Chemeris A.V.
Institute of Biochemistry and Genetics Ufa Science Centre RAS, Ufa, Russia
e-mail: garafutdinovr@mail.ru
** Corresponding author*

Key words: *DNA identification, single nucleotide polymorphism, tetraallelic SNPs, digital encoding, genetic barcode*

Motivation and Aim: Identification of individuals has become an urgent problem of mankind. Recently the DNA identification of personality is actively evolved along with traditional biometric methods [1, 2]. Previously we have developed an approach to identification of individuals by genetic barcoding based on SNPs. The aims of this work are the creation and comparative analysis of genetic barcodes of specific individuals with different ethnicity.

Results: According to approach proposed each possible pair of polymorphic nucleotides in particular SNP is digitized as a unit and received the graphic symbol. Computer software we have written creates genetic barcode in digital and graphic formats. For DNA identification purpose the genetic barcode of each person must be unique. So it is very important to determine the number of SNPs to be analyzed. The feature of our approach is the use of tetraallelic SNPs, which theoretically provide 10^{24} possible combinations of polymorphic nucleotides. We have selected 24 tetraallelic SNPs from SNP database (http://www.ncbi.nlm.nih.gov/snp) and found polymorphic nucleotides by allele-specific PCR. SNP analysis was performed on genomic DNA of more than 100 people – citizens of Russia ethnicities (Russian, Tatar, Bashkir, Yakut, Buryat, Mordovian, Mari, Udmurt, Komi) followed by genetic barcodes creation. For none of the individuals analyzed the genetic barcode does not resemble the barcode of another person despite the fact that all SNPs turned out biallelic. However it is possible for relatives analysis of 24 SNP will be insufficient to provide a uniqueness of genetic barcodes. Therefore we will increase the number of tetraallelic SNPs analyzed up to 48. Earlier it was shown at least 50 biallelic SNPs should be taken for DNA identification [3].

Conclusion: So we have demonstrated the possibility of proposed approach to DNA identification by genetic barcoding based on SNPs. It was found tetraallelic SNPs can provide the uniqueness of genetic barcodes which can be used as a personal identifier of each individual.

Acknowledgements: This work was supported by the Ministry of Education and Science of Russian Federation (contracts No. 14.740.11.1059, 16.518.11.7047).

References:

1. M.A.Jobling, P.Gill. (2004) Encoded evidence: DNA in forensic analysis, *Nat Rev Genet,* 5(**10**): 739-51.
2. R.A.Oorschot, K.N.Ballantyne, R.J.Mitchell. (2010) Forensic trace DNA: a review, *Investig Genet*, 1(**1**): 14.
3. P.Gill. (2001) An assessment of the utility of single nucleotide polymorphisms (SNPs) for forensic purposes, *Int J Legal Med*, 114(**4-5**): 204-10.

"GOLDEN TRIANGLE" FOR FOLDING RATES OF GLOBULAR PROTEINS

Garbuzynskiy S.O.*[1], Ivankov D.N.[1, 2], Bogatyreva N.S.[1], Finkelstein A.V. [1]
[1] Institute of Protein Research, Russian Academy of Sciences, Pushchino, Moscow Region, Russia;
[2] Department of Genome Oriented Bioinformatics, Technische Universitaet Muenchen, Wissenschaftszentrum Weihenstephan, 85354 Freising, Germany
e-mail: sergey@phys.protres.ru
** Corresponding author*

Key words: *protein folding rate; folding time; protein size; protein stability*

Motivation and Aim: An ability of protein chains to form their spatial structures spontaneously is a long-standing puzzle of molecular biology. Experimentally measured rates of spontaneous folding of single-domain globular proteins range from microseconds to hours: the difference (11 orders of magnitude!) is like that between the life spans of a mosquito and the Universe.

Results: We show that physical theory with biological constraints outlines a "golden triangle" limiting the range of folding rates possible for single-domain globular proteins of various size and stability, and that the experimentally measured folding rates fall within this narrow triangle (built without any adjustable parameters), filling it almost completely. In addition, the "golden triangle" predicts the maximal size of protein domains that fold under solely thermodynamic (rather than kinetic) control (this size is about 90 amino acid residues). It predicts also the maximal allowed size of the "foldable" protein domains (it is predicted to be 500 amino acid residues for spherical globules and 600 amino acid residues for oblong or oblate globules); the size of domains found in known protein structures is in a perfect concordance with this limit.

Acknowledgements: This work was supported by Molecular and Cell Biology (#01200957492) program, grants from Howard Hughes Medical Institute (#55005607), RFBR (#10-04-00162a), FASI (#02.740.11.0295), by grant of the Dynasty Foundation and the Russian Young Scientists' grants (#MK-4894.2009.4, #MK-5540.2011.4).

CONTROL OF CULLIN-RING UBIQUITIN LIGASE ACTIVITY BY THE EPSTEIN-BARR VIRUS ENCODED DENEDDYLASE BPLF1

Gastaldello S., Callegari S., Coppotelli G., Hildebrand S., Masucci M.G.*
Department of Cell and Molecular Biology, Karolinska Institutet, S-171 77 Stockholm, Sweden
e-mail: Maria.Masucci@ki.se
** Corresponding author*

Key words: *Epstein-Barr virus, NEDD8, cullin-RING ligase*

The N-terminal domain of the large tegument proteins of herpesviruses is a cysteine protease that promotes efficient virus replication. BPLF1, the Epstein-Barr virus encoded member of this protease family, acts as a deneddylase in infected cells and regulates virus production by modulating the activity of cullin-RING ligases (CRLs). Using a combined bioinformatics, structural and biochemical approach we found that BPLF1 interacts with cullins and stabilizes CRL substrates, which results in the establishment of an S-phase-like cellular environment that is permissive for viral DNA synthesis. We have mapped the site of BPLF1 that interacts with Cul4A to a surface exposed helical domain that is conserved in the homologs encoded by other herpesviruses. The binding site of this BPLF1 domain on the cullin scaffold overlaps with that of the CRL regulator CAND1 and inhibition of CAND1 binding promotes the degradation of deneddylated cullins by the proteasome. The identification of the interacting surface will allow the development of new antivirals.

References

1. Gastaldello S, et al. (2010) A deneddylase encoded by Epstein Barr virus promotes viral DNA replication by regulating the activity of cullin-RING-ligases. *Nature Cell Biol.* 12:351-361.
2. Gastaldello S, et al. (2012) Herpes virus deneddylases interrupt the Cullin-RING ligase neddylation cycle by inhibiting the binding of CAND1. *J Mol Cell Biol.* in press.

EMERGING GENOMIC METHODS AND TECHNOLOGIES

Georgevich G.
BioNanoServe, LLC, Gaithersburg, MD 20878
e-mail: gg@bionanoserve.com, georgevichg@yahoo.com

Motivation and Aim: "Third generation" sequencing technologies are enabling novel applications in genomics, biomarker discovery, diagnostics, systems biology and pharmaceuticals. These technologies use single molecule detection and are based on nanotechnology, nanopores and nano/microfluidics. Testing platforms used in "third generation" sequencing are highly parallel and are capable of extremely high-throughput output. As a result, whole genome sequences can be generated quickly, at costs that are approaching $1,000, and are on a trajectory to reach $100.

Results and Discussion: The use of functional pathways and associations of mutations is becoming practical as the cost/speed of sequencing information improves. We will review some "third generation" systems and the type of data they produce, while addressing the utility of low-cost genomic sequencing to bioinformatics and personalized medicine including its use in gene function discovery and the search for new pharmaceutical targets.

SEARCHING FOR DISTANT HOMOLOGS OF SMALL, NON-CODING RNAs

Giegerich R.
Center of Biotechnology and Faculty of Technology,
Bielefeld University
e-mail: robert@techfak.uni-bielefeld.de

Our appreciation of the functional repertoire of noncoding RNA has increased enormously in recent years. Present RNA-sequencing projects yield a large number of bonafide small RNA transcripts, calling for bioinformatics approaches to filter out promising candidates for experimental study. Potential interaction partners and phylogenetic conservation can provide additional evidence. However at present, there is no fully automated solution either problem.

In the SNAP project (Small Non-coding RNA in Alphaproteo-Bacteria), we identified about 1000 small ncRNA transcripts in *S. meliloti*. Aside from confirmation of small ncRNAs known with *S. meliloti* at that date, search of the Rfam data base with covariance models brought little extra information. 52 trans-encoded RNA transcripts were chosen, for which 39 new family models were build with a mixture of automated and hand-crafting methods.

The talk gives an overview of our findings and then concentrates on two candidate ncRNAs that are widely distributed in the *Rhizobiales*, and cannot be detected with standard methods. Given that we have Internet connection, the talk will end with the audience creating their own search program for finding distant ncRNA homologs.

The SNAP project is joint work with Jan Reinkensmeier, Anke Becker, and Jan-Philip Schl¨uter.

References:

1. Reinkensmeier, Jan and Schl¨uter, Jan-Philip and Giegerich, Robert and Becker, Anke: *Conservation and Occurrence of Trans-Encoded sRNAs in the Rhizobiales* in GENES, Pages: 925-956, 2011.
2. Schl¨uter, Jan-Philip and Reinkensmeier, Jan and Daschkey, Svenja and Evguenieva-Hackenberg, Elena and Janssen, Stefan and J¨anicke, Sebastian and Becker, J¨org and Giegerich, Robert and Becker, Anke: A *genome-wide survey of sRNAs in the symbiotic nitrogen-fixing alpha-proteobacterium Sinorhizobium meliloti* in BMC Genomics, 11(1) , Pages: 245, 2010.

SYSTEMS BIOLOGY ANALYSIS OF COMPLEX DISORDERS

Gilliam C.*[1], Balasubramanian S.[1], Xie B.Q.[1], Sulakhe D.[1], Berrocal E.[1], Maltsev N.[1], Boernigen-Nitsch D.[2], Chitturi C.[3], Paciorkowski A.[4], Dobyns W.[4]
[1] *The University of Chicago, IL, USA;*
[2] *Harvard School of Public Health, Boston, MA, USA;*
[3] *Amrita University, Amritapuri, Kerala, India;*
[4] *University of Washington, Seattle, WA, USA*
e-mail: cgilliam@bsd.uchicago.edu
** Corresponding author*

Key words: *translational medicine, biological networks, gene prioritization*

We present a bioinformatics approach and supporting computational platform - GEDI (http://gedi.ci.uchicago.edu/) - for systems-level analysis of complex heritable disorders such as autism, schizophrenia and diabetes. Our approach is based on a large-scale integration of genomic and clinical data provided by our collaborators, as well as various classes of biological information from over 35 public databases and private collections. This data is used for identification of genes and molecular networks contributing to the phenotypes of interest, as well as for the prediction of additional high-confidence disease genes to be tested experimentally. Our analytical strategy is three-fold and includes (a) the enrichment analysis of high-throughput genomic data (e.g. the results of GWAS and CNV analysis), (b) feature-based gene prioritization and (3) the development of the networks-based disease models for identification of molecular mechanisms involved in pathogenesis of disease of interest. Bayes factor and P-value estimate are used for enrichment analysis; support vector machine algorithm for feature-based prioritization of the candidate genes. Networks-based gene prioritization leverages our previous work [1] and utilizes Heat Kernel diffusion, Random Walk, PageRank with priors, HITS with priors and K-step Markov model algorithms. These algorithms were modified to accommodate variety of weighted data types to be used for gene prioritization (e.g. ranked gene to phenotype associations, weighted canonical pathways data). We will illustrate our approach using analysis of brain connectivity disorders (e.g. agenesis of corpus callosum, autism, schizophrenia) as an example. Our analysis allowed uncovering some of the common molecular mechanisms that underlie these disorders. This knowledge will eventually lead to the development of efficient diagnostic and therapeutic strategies.

[1] Nitsch D, et al. PINTA: a web server for network-based gene prioritization from expression data. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W334-8. Epub 2011 May 20. PubMed PMID: 21602267; PubMed Central PMCID: PMC3125740.

FEATURES ADAPTATION OF CHILDREN OF CHUKOTKA

Godovykh T.V.
Northeast state university, Magadan, Russia
e-mail: tgog@mail.ru

Key words: *adaptation, children, Chukotka, developments, carbohydrate, lipid and albuminous exchanges, transformations*

Motivation and Aim: Growth and development of children in extreme conditions Chukotka occur to the account of a genetic code and factors of an environment of space. Children's population Chukotka is presented nativ and the migrants having various historical degree of adaptation. Studying of mechanisms of adaptation to extreme conditions of the North in development of children will allow to develop methods of the restoration taking into consideration a vector of adapted mechanisms.

Methods and Algorithms: Long-term researches (1996-2005) transformations of a bodyand exchange events of children Chukotka before achievement of 18 years are executed. Indicators of physical development by technics have developed and have accepted at institute of scientific research of anthropology of the Moscow state university, are studied. A lipid, carbohydrates and albumin exchanges, service of calcium and phosphorus of whey of blood (researches are executed in laboratory of biological and inorganic chemistry of department ecological Institute of physiology of natural adaptation, Arkhangelsk).

Results: Transformation bodies during growth and development of children of Chukotka depend on the period of development, a floor, duration of conditions of residing, metisation. Laws and features of transformations of a body on each part of development of children taking into consideration age, population groups, a floor are allocated. Influence of conditions of residing of migrants and metisation nativ on transformations is studied. Exchange processes nativ occur at higher level of energy, since early age to strengthening of humoral regulation, instead of migrants. Seasonal reorganizations of natives are optimum on a metabolism. Adaptation programs nativ have sintoksines elements, metisation shows katatoksines elements. The vector of adaptation of migrants of natives is directed to strengthening sintoksines elements.

Conclusion: In development the system arrives in «effective of development», deducing transformations of a body and metabolic indicators. Programs of adaptation of migrants are directed on a vector of short-term adaptation with katatoksineselements. Programs of adaptation of natives have the long-term period and are understood with prevalence sintoksines elements. Metisation the beginning katatoksines elements. Long residing of migrants (2-3 generations) leads to formation sintoksines elements of the program of adaptation.

Availability: on demand from authors.

THE CENTRAL REGULATORY CIRCUIT OF THE MACROCHAETE MORPHOGENESIS GENE NETWORK: A MODEL OF FUNCTIONING

Golubyatnikov V.P.*[1, 3], Bukharina T.A.[2], Furman D.P.[2, 3]
[1] *Sobolev Institute of Mathematics, SB RAS, Novosibirsk, Russia;*
[2] *Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia;*
[3] *Novosibirsk State University, Novosibirsk, Russia*
e-mail: glbtn@math.nsc.ru
** Corresponding author*

Key words: *mechanoreceptor, AS-C, central regulatory circuit, model*

Motivation and Aim: The key element in the gene networks providing the morphogenesis of drosophila mechanoreceptors is the *achaete-scute* gene complex (*AS-C*). The expression of *AS-C* is controlled by the central regulatory circuit (CRC) [1], comprising eight genes. The proteins they code for implement the induction–repression relations between the CRC components and provide for a certain level of the AS-C protein in the mechanoreceptor precursor cell. The goal of the work was to construct a mathematical model describing the CRC functioning in this cell.

Methods and Algorithms: The constructed model describes the CRC functioning as a dynamic system with a set of positive and negative feedbacks.

Results: Activity regulation of the *AS-C* genes under the CRC control in a dimensionless form is described by the following system of equations:

$$\frac{dx_1}{dt} = \frac{S_1(D\cdot x_1) + S_2(x_3) + S_6(x_5)}{(1+G\cdot x_2)(1+E\cdot x_1)} - x_1;\ \frac{dx_2}{dt} = \frac{C_2}{1+x_4} - x_2;\ \frac{dx_j}{dt} = S_j(D\cdot x_1) - x_j,\ j = 3,4,5.$$

The variables x_1, x_2, x_3, x_4, x_5 denote the concentrations of AS-C, Hairy, Senseless (SENS), Scratch (SCRT), and Charlatan (CHN) proteins, respectively, and parameters D, G, and E are the concentrations of Daughterless (DA), Groucho (GRO), and Extramacrochaete (EMC) cofactor proteins. The effect on *AS-C* via positive feedbacks is described by the sigmoid functions $S_i(X)$: i=1,2,...,6. The products ($D\cdot x_1$), ($G\cdot x_2$), and ($E\cdot x_1$) correspond to the heterodimers AS-C/DA, HAIRY/GRO, and AS-C/EMC.

Conclusions: Analysis of the constructed model suggests several conclusions: (1) the system does not reach the mode of infinite increase in the content of AS-C proteins in the cell at any sets of parameters; (2) the content of AS-C proteins increases in a stepwise manner, which is likely to reflect successively switched-on activating impacts via SENS, CHN, and SCRT; and (3) the CRC functioning does not follow a cyclic pattern.

Acknowledgements: Supported by the RFBR grants nos. 12-01-00074, and 09-04-12209-ofi_m; SB RAS (integration projects nos. 80 and 136), RAS (Program A.II.6 and project no. VI.45.1.3), grant no. NSh-5278.2012.4, and state contracts nos. 16.512.11.2129 and 16.522.12.2006.

References:

1. D.P. Furman, T.A. Bukharina, (2009) The Gene Network Determining Development of *D. melanogaster* Mechanoreceptors, *Comp. Biol. Chem.*, **33:** 231–234.

AN INVERSE PROBLEM OF IDENTIFICATION OF PARAMETERS IN ONE GENE NETWORK MODEL

Golubyatnikov I.V.
Sobolev Institute of Mathematics, SB RAS, Novosibirsk, Russia
e-mail: ivan.golubyatnikov@gmail.com

Key words: *inverse problems, a Central Regulatory Contour model*

Motivation and Aim: Inverse problems appear on different levels of gene network studies: in reconstruction of gene network structure, in determination of coefficients in the models etc., see [1] for detailed exposition.

Methods and Algorithms: We continue our studies of a model of one Central Regulatory Contour (CRC) constructed in [2,3].

Results: Following [2,3], we describe regulation of *achaete_scute* complex in CRC by the nonlinear dynamic system:

$$\frac{dx_1}{dt}=\frac{S_1(D\cdot x_1)+S_2(x_3)+S_6(x_5)}{(m_2+G\cdot x_2)(m_1+E\cdot x_1)}-k_1x_1;\ \frac{dx_2}{dt}=\frac{C_2}{1+B_2x_4}-k_2x_2;\ \frac{dx_j}{dt}=S_j(D\cdot x_1)-k_jx_j,\ j=3,4,5. \quad (1)$$

In contrast with [3], this system does not have a dimensionless form. Positive coefficients m_1, m_2, A_j, B_j,C_2, k_i, j=1,…6; i=1,…5 in these equations are assumed to be unknown. The variables x_1, x_2, x_3, x_4, x_5 denote concentrations of AS-C, HAIRY, SENS, SCRT, CHN, respectively; and D, G, E are concentrations of DA, GRO, EMC, considered as parameters of the dynamical system (1), see [2]. As in [3], the sigmoid functions

$S_i(X)=\frac{A_i\cdot X}{B_i+X}$, $i=1,2,...,6,$ describe positive feedbacks in CRC, and $D\cdot x_1$, $G\cdot x_2$, $E\cdot x_1$ correspond to the heterodimers DA–(AS-C), GRO–HAIRY, EMC–(AS-C).

We show that under some natural physical assumptions the inverse problem of identification of these parameters has a unique positive solution $\{m_1, m_2, A_j, B_j, C_2, k_i\}$.

Acknowledgments: The work was supported by RFBR grant 12-01-00074.

References:

1. Computational Systems Biology, (2008) (Ed.:N.A.Kolchanov), SB RAS.
2. D.P.Furman, T.A. Bukharina, (2009) The Gene Network Determining Development of
3. *D. melanogaster* Mechanoreceptors, *Comp. Biol. Chem.*, **33:** 231–234.
4. T.A.Bukharina, et al (2012) Model Investigation of Central Regulatory Contour of Gene Net of *D. melanogaster* Macrochaete Morphogenesis, *Russian Journal of Developmental Biology,* **43:** 49–53.

DEVELOPMENT OF A NOVEL PYROSEQUENCING-BASED METHOD FOR STUDYING *E. COLI* DIVERSITY AND MICROBIAL SOURCE TRACKING

Goodman A.*, Montana A., Neal E., VanderKelen J., Black M., Kitts C., Dekhtyar A.
California Polytechnic State University, San Luis Obispo, CA, USA
e-mail: agoodman@calpoly.edu
** Corresponding author*

Key words: *microbial genome diversity, pyrosequencing, assay development, modeling*

Motivation and Aim: Fecal contamination of food and water supply frequently cause public health problems around the world. There is an urgent need for rapid and inexpensive method of identifying the sources of contamination. The presence of *E. coli* is typically used as an indicator of fecal contamination; however, identifying the source of specific strains of *E. coli* remains a major challenge. We have developed a novel method for the identification of *E. coli* strains by generating molecular fingerprints via simultaneous, multi-locus pyrosequencing of the ribosomal RNA (rRNA) operon.

Methods and Algorithms: Previously, rRNA genes have been eliminated from the list of potential targets in sequence-based assay development because rRNA operons are present in seven copies in each *E. coli* genome [1]. Our method takes advantage of the multiple copies of the rRNA operon to help discriminate between closely related strains. We designed a novel assay that uses PCR to amplify all seven copies of the ribosomal RNA intergenic regions and then sequences them together in a single reaction. The two polymorphic intergenic transcribed spacer (ITS) regions that reside between the rRNA segments (16S, 23S, and 5S) are used as targets to distinguish between similar strains. The raw pyrosequencing data from each reaction is a pattern of peaks. While these data cannot be used for sequence analysis due to the use of multiple templates, they may be used to differentiate between strains. These patterns are reproducible and characteristic of each strain, resulting in output that is analogous to a fingerprint; therefore, we refer to the patterns as pyroprints. By pyrosequencing multiple templates that differ in their sequences, the effect of single-nucleotide polymorphisms (SNPs) may be amplified through a "ripple effect": a difference at one of the seven loci resulting in changes of multiple subsequent signal peaks. Whether or not the "ripple effect" is observed depends on the genomic sequence and the dispensation order. To optimize assay parameters, we developed a pyroprint modeling program to model pyroprints from 38 finished *E. coli* genomes. We use Pearson correlation coefficient to compare pyroprints.

Results: Pyroprints are highly reproducible (>99%). Our collection currently contains nearly 3,000 pyroprints from *E. coli* isolated from 16 avian and mammalian hosts. Some strains appear to be shared by multiple hosts, while others appear host-specific. We are developing web based pyroprint database and analytical tools and extending this method to other bacterial species.

Conclusion: Pyroprinting is a novel method for identification of *E. coli* strains. Pyroprinting has the following advantages: simple protocol, reproducible, low cost, easy transfer from lab to lab, easy to scale up, and good discrimination between closely related strains.

References:

1. K.M. Ivanetich et al. (2006) Microbial source tracking by DNA sequence analysis of the Escherichia coli malate dehydrogenase gene. *J Microbiol Methods.* 67(3):507-26.

INTEGRATION OF – OMICS

Govorun V.M.
Research Institute of Physico Chemical Medicine, Moscow, Russia

Recently rapid development allowed for explosion in amount of data and sensitivity of detection methods in Genomics, Transcriptomics, Proteomics and Metabolomics. However while the interdependence of – Omics *in vivo* is of no doubt, integration of experimental data mostly results in unpreceded complexity – whereas analysis and further predictions seem almost impossible.

Promising is the approach of Omics integration in bacterial studies, ta decade ago bacterial cells seemed to be tremendously simple compared to eukaryotic ones. The range of studies aimed at bacterial analysis however demonstrated comparable complexity in Omics result in bacteria, where the difference is only in sizes of genomes and numbers of variable proteins, RNAs and metabolites.

We have developed several experimental and bioinformatics pipelines allowing for data integration in experimental results. We demonstrate the emergence effects appearing from integration of Omics for several different organisms: Mollicutes including spiroplasma, mycoplasma and acholeplasma, H. pylory. Further we demonstrate the scalability of the approaches developed for data integration in human microbiome analysis, and variety of eukaryotic organisms.

STRUCTURAL AND FUNCTIONAL PROTEOMICS OF THE HUMAN PROTEIN SYNTHESIZING SYSTEM

Graifer D.M.*[1], Bulygin K.N.[1], Khairulina Yu.S.[1], Sharifulin D.E.[1], Ven'yaminova A.G.[1], Frolova L.Yu.[2], Karpova G.G.[1]

[1] *Institute of Chemical Biology and Fundamental Medicine SB RAS, Novosibirsk, Russia;*
[2] *Engelhardt Institute of Molecular Biology RAS, Moscow, Russia*
e-mail: graifer@niboch.nsc.ru
** Corresponding author*

Key words: *eukaryotic ribosome; decoding site; mRNA analogues; photoaffinity cross-linking; eukaryotic ribosomal protein S15; stop codon recognition; eRF1*

Protein synthesis is one of the basic events of the cell life. It takes place on ribosomes, very complicated cellular ribonucleoprotein machineries translating genetic information incoming as mRNA, and involves a number of assistant proteins (translation factors). Specific interactions of proteins during translation process underlie the work of the ribosomal machinery, and knowledge of the structural basis of these interactions is of principal importance for life sciences. These interactions in prokaryotes are known at the atomic level due to X-ray crystallography. However, structural and functional proteomics of the human protein synthesizing system is much less studied since ribosomes from higher organisms were not yet crystallized for the X-ray analysis. Here, we applied a site-directed cross-linking approach to study fine structure of mRNA binding site of the human ribosome and of a site of stop codon recognition in translation termination factor eRF1. Using a set of labeled mRNA analogues bearing cross-linker at designed locations, we determined peptides of ribosomal proteins involved in the formation of the mRNA binding site, and peptides of eRF1 recognizing mRNA stop codons. We found that eukaryote-specific peptide of ribosomal protein S26e is a key player in accommodation of mRNA region 5' of the codons interacting with tRNAs suggesting that it is involved also in the interaction with eukaryote-specific initiation factor eIF3 [1]. Eukaryote/archaea-specific decapeptide of ribosomal protein S15e was revealed at the ribosomal decoding site and most probably interacts with eRF1 [2]. We clarified the keystone aspect of protein synthesis termination related to the recognition of stop codon purines by eRF1. We found out that A and G are recognized by different N domain conformations of eRF1, which provides its ability to recognize all three stop codons, and discovered that the universally conserved dipeptide 31-GT-32 is the key player in this process [3,4]. The data obtained provide new insights into molecular basis of mammalian protein synthesis and are of great importance for understanding the nature of pathologies related to disturbances of any step of this process.

This study was supported by RFBR grant # 11-04-00597 to G.K. and by the Russian Academy of Sciences Presidium program "Molecular and cell biology" (grant to G.K.).

References:

1. Sharifulin D. et al. (2012) Nucleic Acids Res. In press (doi: 10.1093/nar/GKR1212).
2. Khairulina J. et al. (2010) Biochimie 92:820-825.
3. Bulygin K.N. et al. (2010) RNA 16:1902–191.
4. Bulygin K.N. et al. (2011) Nucleic Acids Res. 39:7134–7146.

COMPLEX COMPUTATIONS AND WORKFLOWS IN MOLECULAR BIOLOGY

Grekhov G.*, Fursov M.Y., Kandrov D.
Novosibirsk Center of Information Technologies 'UniPro'
e-mail: ggrekhov@unipro.ru
** Corresponding author*

Key words: *UGENE, Workflow Designer, computational schemes*

Motivation and Aim: Bioinformatics methods play the dominant role in system biology today. During the past decade there was developed number of unique methods in bioinformatics that played vital roles in technological and scientific breakthroughs. Integrated bioinformatics solutions that combine different statistics, algorithms, automated methods of data retrieval in areas of genomics, transcriptomics and proteomics becomes more valuable today than never before.

Such systems allow to researcher focusing on an experiment and use more complex computational schemas than it was possible when using separate programs and datasets.

To join different bioinformatics methods into a complex computational workflow there is a need in a new domain specific language that will be easy to use and powerful enough at the same time.

Results, Conclusion and Availability: We have designed and implemented an interactive software solution called Workflow Designer. The key idea of Workflow Designer is to make the process of routine tasks automation as simple as possible and make it available and understood by non-programmers.

A user of Workflow Designer draws a computational scheme from predefined elements (computational blocks). Each block contains a complete textual description of what will be done during its execution. The description is automatically adapted to the parameters and environment this computational block is used in. The final workflow schemas comprise reproducible, reusable and self-documented research routines, with a simple and unambiguous visual representation suitable for publications.

Workflow Designer is a part of UGENE[1][2] genome analysis suite and supports all computational methods and data retrieval options available in UGENE. Additionally you can create custom workflow elements.

A ready to use version of the software as well as the complete source code is freely available under the GPL license from the UGENE web site or as a part of major Linux distributions. The solution is available for Linux, Windows and MacOS X platforms.

References:

1. K. Okonechnikov, O. Golosova, M. Fursov. (2012) Unipro UGENE: a unified bioinformatics toolkit, Bioinformatics 2012: bts091v1-bts091.
2. UGENE web site: http://ugene.unipro.ru

SEARCH FOR FUNCTIONAL PATHWAYS FOR INTRAMEMBRANE ASPARTIC PROTEASE *IMPAS1/SPP*

Grigorenko A.P.[1, 2], Moliaka Y.[1], Alexandrov I.[1], Rogaev E.I.[1, 2]
[1]*University of Massachusetts Medical School, USA;*
[2]*Vavilov Institute of General Genetics, Research Center of Mental Health, Moscow*
Correspondence should be addressed to Evgeny.Rogaev@umassmed.edu

Motivation and Aim: Previously, we and others have identified five genes in human genome for novel polytopic family of intramembrane aspartic proteases IMPAS/IMP (or H13, SPP/SPPL or PSH) homologous to presenilins (Grigorenko et al, 2002; Moliaka et al, 2004; Grigorenko et al, 2004). This type or proteases supposed to be critical for cleavage of type II transmembrane proteins (including signal peptides) in transmembrane domain. It has been predicted that IMPs substrates are type II membrane proteins with C-terminal part oriented into the lumen. *In vitro* assay demonstrated that SPP/IMP1 cleaves short signal peptide remnants tethered in ER membranes. This activity may generate short signal sequence that is essential for HLA-E epitope (Weihofen et al, 2002). However, the major functions *in vivo* and proteolytic substrates of IMP1 proteins *in vivo* are unknown.

Methods and Algorithms. We generated the knockout mouse model and cellular models (*mIMP1-/-)* for *IMPAS1/SPP* gene. Several approaches were taken to elucidate molecular pathways regulated by IMP1 gene: testing of apoptosis and autophagy events; comparative transcriptome profiles of all protein-encoding genes in *IMP1* knockout and control mouse brains; real-time PCR gene expression study and signaling pathway analysis with Cignal™ Reporter Assays in *IMP1-/-* and *IMP1-/+* or *IMP1+/+* cells.

Results. We have shown that *mIMP1* gene is crucial for early embryonic development and plays an important role in brain development (unpublished). We have identified alteration of certain transcriptional factors in *mIMP1* knockout mice cells suggesting that IMP1/SPP is essential for signal transduction regulation in early development. Empirically, we have excluded the number of type II protein receptors as putative substrates for IMP1 protease but confirmed the cleavage of HCV (viral) core protein and showed that both human IMP1/SPP and its *C.elegans* ortholog protease capable to cleave (in co-transfected cells) the C-terminal transmembrane domain of multipass transmembrane protein presenilin 1. The endogenous substrates for IMP1/SPP with single or multiple transmembrane domains involved in signaling regulation in development can be predicted and have to be identified.

Acknowledgements. Supported by NIH/NIA AG029360.

References:

1. Grigorenko AP, Moliaka YK, Korovaitseva GI, Rogaev EI. Biochemistry (Mosc) 2002; 67: 826-35.
2. Grigorenko AP, Moliaka YK, Soto MC, Mello CC, Rogaev EI. Proc Natl Acad Sci U S A 2004; 101: 14955-60.
3. Moliaka Y., Grigorenko A., Madera D., Rogaev E. FEBS Letters (2004); **557**:185-192. 4.
4. Weihofen A et al, Science 2002; 296: 2215-8.

MOLECULAR EVOLUTION OF HUMAN PROTEIN-CODING GENES IN THE LIGHT OF BRAIN ORGANIZATION

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: *human brain, positive selection, neutral evolution, protein-coding genes*

Motivation and Aim: There are several well known databases such as SELECTOME [1] or Human PAML browser [2] deposited the information about molecular evolution mode of genes controlling development and functioning of human brain. In these databases the evolution mode of protein-coding genes is determined by examining the ratio of the rates of nonsynonymous to synonymous nucleotide substitution (Ka/Ks) using PAML program package and branch-site methods. However there are clear evidences that the branch-site methods often generate false positives results [3]. With this in mind we recalculated the evolution modes of all known protein coding genes using various Ka/Ks calculation methods on the basis of pairwise sequence comparison without differentiation of sites in gene alignment. Using this recalculated data we reanalyze the structures of human brain in which positively selected and neutrally evolved genes are expressed.

Methods and Algorithms: The evolution modes of protein coding genes were inferred using 15 Ka/Ks calculation methods implemented in the KAKS_Calculator 2.0 program package [4]. Pairwise alignments of protein coding genes represented the main evolutionary steps of hominids (*Homo sapiens*, *Pan troglodytes, Gorilla gorilla, Pongo pygmaeus*) were extracted from ENSEMBL Rel. 62. The identification of human brain tissues in which positively selected genes are expressed was based on the Allen Human Brain Atlas [5]. To find the genes that highly expressed in human brain we use strict threshold of ≥ 9 [5].

Results: We focused our attention on genes which are highly expressed in human brain and positively selected on different hominids tree branches. It was shown that the most of such genes on the phylogenetic tree branch passed from Homo-Pan ancestor to modern human expressed in frontal and temporal gyri (cerebral cortex). It is of interest that this trend of positively selected genes expression occurred in all tree branches with exception of the branch from Gorilla ancestor to Homo-Pan ancestor which possessed zero number of positively selected genes.

Conclusion: The global trend for positively selected genes expression in different branches of hominids phylogenetic tree was found. This trend is the expression of these genes in frontal and temporal parts of cerebral cortex.

Acknowledgement: The work supported by SB RAS project 136; RAS project 6.8.

References:

1. Proux E., et al. (2009) Selectome, *Nucleic Acids Res,* **37:** D404-407.
2. Nickel G.C., et al. (2008) Human PAML browser, *Nucleic Acids Res,* **36:** D800-D808.
3. Nozawa M., et al. (2009) Reliabilities of identifying positive selection by the branch-site and the site-prediction methods, *PNAS USA*, **106**: 6700-6705.
4. Wang D., et al. (2010) KaKs_Calculator 2.0, *Genomics Proteomics Bioinformatics*, **8:**77-80.
5. Allen Human Brain Atlas, http://www.brain-map.org/

IMPORTANT ROLE OF THE miRNA CHANGES IN THE *HOMO NEANDERTHALENSIS* AND *HOMO DENISOVA* EVOLUTION

Gunbin K.V.*[1], Afonnikov D.A.[1], Kolchanov N.A.[1], Derevyanko A.P. [2]
[1] *Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia, e-mail: genkvg@bionet.nsc.ru;*
[2] *Institute of Archeology and Ethnography, SB RAS, Novosibirsk, Russia*
** Corresponding author*

Key words: *Homo neanderthalensis, Homo denisova, miRNA, molecular evolution*

Motivation and Aim: In 2010, the genomes of the *H. neanderthalensis* (*H. n.*) [1] and *H. denisova* (*H. d.*) [2] were sequenced. At the present time it becomes obvious that the most rapid evolutionary transformations, observed at the earliest stages of *H. sapiens sapiens* (*H. s.s.*) divergence, have to do, first of all, with the change in the miRNA genes [3]. However, the question about the molecular-genetic changes, which led to the emergence of extinct archaic Neandertal and Denisovan humans, remains open. With this in mind, we have conducted a comparative computer analysis of the *H. s.s.* miRNA genes with those of ancient humans (*H. n.* and *H. d.*).

Methods and Algorithms: In our analysis the evolutionary conservative miRNA *H. s.s.* genes were taken from ENSEMBL database, *H. n.* and *H. d.* orthologous of these genes were taken from UCSC ftp storage. Thorough selection procedures were applied in order to choose functionally important *H. n.* and *H. d.* miRNA with evolutionary meaningful changes from *H. s.s.* orthologs. After that the functional annotation of these selected miRNA was done using experimental and theoretical information about miRNA/mRNA interaction deposited in starBase [4]. At the final step, we performed the functional enrichment (permutation) test comparing the occurrence rates of functional annotations of target genes for rapidly evolving *H. n.* and *H. d.* miRNAs with those of target genes for all sequenced *H. n.* and *H. d.* miRNAs.

Results and Conclusion: Based on the study of the functions of miRNA-target genes for rapidly evolving *H. n.* and *H. d.* miRNAs and changes in the secondary structure of these miRNA precursors, it was shown that changes in miRNA genes could have a prominent role in the *H. n.* and *H. d.* evolution, namely in the development and functioning of its brains. For example, it was shown that the main evolutionary changes in the expression of miRNA target genes for rapidly evolving *H. d.* miRNAs may be related with the prefrontal cortex which is responsible for thinking and talking. The evolutionary changes in *H. n.* miRNAs mainly related with cerebellum and hindbrain which is consistent with the current data on *H. n.* anatomy and life.

Acknowledgements: The work supported by RFBR (11-06-12006-ofi-m-2011), SB RAS (projects 130, 39, 93), RAS (project 6.8 and program 28), State contract 82/201 and Scientific school 5278.2012.4.

References:

1. Green R.E., et al. (2010) A draft sequence of the Neandertal genome. *Science,* **328:**710-722.
2. Reich D., et al. (2010) Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature,* **468**:1053-1060.
3. Somel M., et al. (2011) MicroRNA-driven developmental remodeling in the brain distinguishes humans from other primates. *PLoS Biol.,* **9**:e1001214.
4. Yang J.H., et al. (2011) starBase: a database for exploring microRNA-mRNA interaction maps from Argonaute CLIP-Seq and Degradome-Seq data. *Nucleic Acids Res*., **39**:D202-D209.

HIGHWAYS IN THE HORIZONTAL TRANSFER OF EUBACTERIAL Fpg AND Nei GENES

Gunbin K.V.*[1], Afonnikov D.A.[1], Zharkov D.O.[2]
[1] *Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia;*
[2] *Institute of Chemical Biology and Fundamental Medicine, SB RAS, Novosibirsk, Russia*
e-mail: genkvg@bionet.nsc.ru
** Corresponding author*

Key words: *Fpg_Nei protein family, protein structure evolution, horizontal gene transfer*

Motivation and Aim: The Fpg_Nei protein family encodes well known base-excision repair enzymes found in eukaryotes and prokaryotes. In 2011, a thorough analysis of the Fpg_Nei protein family structural motifs evolution has been made [1]. As a result of this analysis 7 latent protein structural characters of Fpg_Nei protein family were discovered [1]. However, the impact of horizontal gene transfer (HGT) to the evolution of these proteins was not analyzed in [1]. Thus the aim of this work is to estimate the influence of the HGT events on the Fpg_Nei protein evolution.

Methods and Algorithms: Protein sequences belonging to Fpg_Nei family were retrieved from RefSeq rel. 46. We reconstructed the set of consensus phylogenetic Fpg_Nei subtrees using PhyloBayes 3.3 program package [2] on the basis of best fitted models of protein evolution. Fitting of different protein evolution models to the Fpg_Nei family subsamples were done using ProtTest 3.0 program [3]. It is of importance that all of reconstructed subtrees possessed unresolved nodes but root nodes of these subtrees are strictly statistically supported (Bayesian posterior probability ≥ 0.99). Thus the potential HGT events should be recovered only in these subtrees. The examination of HGT routes in subtrees were done using SplitsTree 4 program [4].

Results and Conclusion: We focused our attention on Fpg_Nei subtrees encompassing Prokaryota superclade. It was shown that the vast majority of HGT events concentrated in Fpg clade, even including potential HGT events between pro- end eukaryotes. Nei clade is characterized by the relative low rate of HGT events in prokaryotes only. Therefore it is likely that Fpg-specific latent protein structural characters discovered in [1] reshuffled during prokaryotes (and, probably, eukaryotes) evolution several times. This possibility complements and complicates the picture of Fpg-specific latent protein structural characters evolution described in [1]. Thus our results show the necessity in more detailed study of Fpg-subfamily structural evolution based on HGT events consideration.

Acknowledgements: The work supported by RFBR (11-04-01771-a), SB RAS (projects 130, 39), RAS (project 6.8 and program 28) and Scientific school 5278.2012.4.

Availability: Results available upon request.

References:

1. Barrantes-Reynolds R., et al. (2011) Using shifts in amino acid frequency and substitution rate to identify latent structural characters in base-excision repair enzymes, *PLoS One,* **6:** e25246.
2. Lartillot N., et al. (2009) PhyloBayes 3, *Bioinformatics,* **25:** 2286-2288.
3. Darriba D., et al. (2011) ProtTest 3, *Bioinformatics,* **27:** 1164-1165.
4. Huson D.H., Bryant D. (2006) Application of Phylogenetic Networks in Evolutionary Studies, *Mol Biol Evol.*, **23**: 254-267.

PEFF DB: THE WEB-AVAILABLE DATABASE OF PROTEIN EVOLUTIONAL AND FUNCTIONAL FEATURES

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

Key words: *orthologous protein families, gene networks, molecular evolution*

Motivation and Aim: In 2008–2009, great doubts about validity of the nonsynonymous to synonymous nucleotide substitutions rate ratio as a robust criterion of positive selection were raised [1]. Alternative approaches to study the molecular evolution modes of proteins based on the rate of change (V_P) of various properties of amino acids in the course of protein evolution and allows analyzing the evolutionary mode even at the deep inner tree branches. We improve the V_P analysis by implementing the permutation test for comparison of the simulated molecular evolution of sequences from the protein family with real evolution of these proteins [2]. The aim of this work was to make a massive computer analysis of all available well-corroborated orthologous protein groups (OPGs) using new methodology and represents the results of this analysis as a web-available database.

Methods and Algorithms: The vertebrate and invertebrate OPGs were taken from MetaPhOrs database [3]. The SAMEM [4] pipeline of protein molecular evolutionary analysis was used. The analysis of the evolutionary modes of proteins at each branch of the vertebrate and invertebrate phylogenetic trees was made using approach that allows analyzing evolutionary mode even at the deep inner tree branches [2]. We linked our data on molecular evolution modes with 21 internet-available databases deposited information about protein domains, structure, function, position of protein in gene network and chemical kinetic data.

Results: Using all data deposited in PEFF database it was shown that the Vertebrate and Invertebrate internal tree branches enriched with statistically rare amino acid replacements strictly correspond to evolutionary aromorphoses. For example, these branches marked: 1) the adaptation of Vertebrates to terrestrial environments, 2) the origin of Amniota, 3) the divergence of primitive and placental mammals, 4) the divergence of Insecta and Diptera. In the developed database we also conducted the functional enrichment permutation test of OPGs containing statistically rare amino acid replacements at each inner tree branch. This test allows us to uncover various features of Metazoan gene networks evolution.

Conclusion: An internet-available database containing information about the molecular evolution of OPGs, integrated with data on their structure and function was made.

Acknowledgements: The work supported by RFBR (11-04-01771-a), SB RAS (projects 130, 39), RAS (project 6.8 and program 28) and Scientific school 5278.2012.4.

Availability: http://pixie.bionet.nsc.ru/peff/

References:

1. Drummond D.A., Wilke C.O. (2009) The evolutionary consequences of erroneous protein synthesis. *Nat Rev Genet*., **10**:715-724.
2. Gunbin K.V. et al. (2011) Molecular evolution of cyclin proteins in animals and fungi. *BMC Evol Biol*., **11:** 224.
3. Pryszcz L.P. et al. (2011) MetaPhOrs, *Nucleic Acids Res.,* **39**: e32.
4. SAMEM (2009-2012), http://pixie.bionet.nsc.ru/samem/

COMPUTER ASSISTED STUDY OF THE GTF2I PROTEIN REPEATS EVOLUTION

Gunbin K.V. *[1], Ruvinsky A.O.[2], Afonnikov D.A.[1]
[1] *Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia;*
[2] *The Institute for Genetics and Bioinformatics, University of New England, Armidale, NSW 2351, Australia*
e-mail: genkvg@bionet.nsc.ru
** Corresponding author*

Key words: *GTF2I protein family, domain function, protein structure evolution*

Motivation and Aim: The GTF2I gene family consists of three subfamilies (*gtf2i*, *gtf2ird1* and *gtf2ird2*) of genes encoding transcriptional factors [1]. The distinguishing feature of this gene family is the presence of repeats in the primary stricture of encoded proteins. It is of interest that the repeats composition is the major factor determining the molecular function of these proteins [1]. The aim of this work was to analyze the molecular evolution of these repeats and to predict their structure-functional characteristics.

Methods and Algorithms: Protein sequences of the GTF2I family retrieved from ENSEMBL rel. 63 and KEGG Orthology rel. 59.0. We aligned proteins by PROMALS3D and extracted from the whole alignment the repeat alignments. The repeat consensus phylogenetic trees reconstructed using PhyloBayes 3.2f and MrBayes 3.2.1 programs. On the basis of the obtained sets of Bayesian binary trees we made the phylogenetic networks (via Dendroscope 3) and analyze the possibilities of repeats recombination. Using repeats structures deposited in PDB we divided repeats to sets of secondary structures elements (SSE). The analysis of the evolutionary modes of repeats and its particular SSE has been made using approach described in [2]. Moreover, we analyzed the trends of physicochemical changes in their evolution using nonparametric statistics which allow us to predict the structural changes in the globule in GTF2I protein family evolution.

Results and Conclusion: The unique mutational pattern for each GTF2I repeat (for each site and secondary structure) was found. It was shown that this pattern could be caused by repeats functional specialization (for example, DNA-binding [1]). It was shown that the statistically rare types of amino acid replacements in repeats occurred during the emergence of tetrapods, mammalians and birds. This fact elucidates the possible role of GTF2I proteins in adaptation of vertebrates to different environments. At last we describe the most probable positions and function of each repeat in protein globule and their changes in evolution on the basis of thorough analyzing of the evolutionary changes of all known amino acid physicochemical properties. Thus, on the basis of comprehensive evolutionary analysis of GTF2I protein domains we found that this protein family has an important role in vertebrate evolution.

Acknowledgements: The work supported by RFBR (11-04-01771-a), SB RAS (projects 130, 39), RAS (project 6.8 and program 28) and Scientific school 5278.2012.4.

References:

1. Palmer S.J. et al. (2010) Negative autoregulation of GTF2IRD1 in Williams-Beuren syndrome via a novel DNA binding mechanism, *J Biol Chem.* **285**:4715-4724.
2. Gunbin K.V. et al. (2011) Molecular evolution of cyclin proteins in animals and fungi. *BMC Evol Biol*., **11:** 224.

MODELING EMERGENT PROPERTIES OF BIOLOGICAL SYSTEMS WITH AN AGENT-BASED SIMULATION SUITE

Henderson R.
National Institutes of Health, Bethesda, Maryland USA
e-mail: rh@nih.gov

Key words: *agent-based modeling, systems biology, emergent properties, genetic regulatory networks, metabolic regulatory networks, multi-agent simulation*

Motivation and Aim: Living systems are complex systems. As such, they have emergent behaviors: input-response properties that can be observed but not predicted by first order knowledge of the functions of the system's components. Understanding emergent behaviors of complex biological systems requires modeling and simulation of large and detailed prototypes. Models must be both expressive and scalable to capture the size and complexity of molecular and cellular networks. Existing genetic regulatory pathways analysis tools allow researchers to map their experimental data onto gene networks, but do not allow researchers to actively simulate dynamic perturbations to specific nodes. For example, existing tools allow researchers to see that repression of A activates B, which then activates C and D, and then activation of D represses E. However, there is no way for researchers to easily see the effect on E by the activation of A without manually walking through the network node by node. In this example, only five nodes are represented; the problem becomes increasingly complex, however, since the solution space increases exponentially as the number of network components increases linearly. Consequently, for realistic biochemical networks with hundreds or thousands of components, existing tools offer no mechanism with which researchers can readily study metabolic or genetic effects throughout the network and at a distance from an inhibitory or excitatory site.

Methods and Algorithms: In this report we present a modeling and simulation approach, GRANITE, (Genetic Regulatory Analysis of Networks Investigational Tools Environment), an agent-based modeling and multi-agent simulation approach to modeling large, complex, and dynamic systems. We show that GRANITE is expressive enough to capture any kind of interaction network, can modularly use any kinetic model, is computationally tractable and scalable, and allows researchers to interact and dynamically perturb the system at different hierarchical levels to learn its rules for emergent behavior.

Results: We have demonstrated the GRANITE capability on metabolic networks: specifically the mycolic acid biosynthesis pathway of the *Mycobacterium tuberculosis*. The agent-based model has been compared to Flux Balance Analysis (FBA) and shown to be able to emulate the internal and external properties of the system as modeled by FBA.

Conclusion: We show that the approach is scalable and computationally efficient to allow researcher interaction with a dynamically evolving simulation. The GRANITE tool enables the researcher to propose and test systems-level hypotheses and make predictions for laboratory experiments to validate or refute these hypotheses.

Availability: The source code for GRANITE is freely available for MS Windows, Linux, and Mac OS X at http://exon.niaid.nih.gov/data/granite.zip

MODELING ASPECTS OF THE "VIRTUAL CELL"

Hofestädt R.
Bielefeld University
e-mail: hofestae@techfak.uni-bielefeld.de

During the last decades molecular genetics could identify fundamental mechanisms and components of the cell system. Behind the omics data detection and analysis also the discussion of the DNA as a programming language became relevant. Furthermore different Virtual Cell projects started – Ecell is one of these famous projects only. However, C. Venter started the idea of Synthetic Biology – the idea to construct new bacteria – new life. Synthetic Biology can only be successful if the Virtual Cell project will be successful. Therefore, we started the Cellmicrocosom (www.cellmicrocosmos.org/) project. This new tool allows the editing of real and artificial cells and is already used in schools for education reasons. Behind the education application we use this tool for the design of membrane and for the 3D representation of biological networks (talk of B. Sommer will present this project). Regarding the biological networks of the cell we have to discuss the useful method for modelling and simulation and the analysis of these data by analysis algorithms (talk of David Braun). However, the kernel of a Virtual Cell simulator is the simulation of biochemical processes. Until now it is still an open question which kind of formalization we have to use in practice to solve this question.

Bjoern Sommer will present in his talk the Cellmicrocosmos project and he will show how we can edit real and artificial cells (regarding cell components). Until now this tool is not able to simulate metabolic processes. Based on this static virtual cell concept this talk will discuss and motivate the modelling and simulation process of biological networks. For modelling of biological networks we use graph theoretical aspects and implemented the tool VANESA to edit, create, extend and analyse biological networks (http://vanesa.sourceforge.net/). For simulation of these processes we use Petri Net application and implemented a new Petri Net simulation tool, which is already embedded into the VANESA system.

References

1. R. Hofestädt, Fundamental Features of Metabolic Computing. Proceedings of MICAI 2011, LNCS (Advances in Soft Computing) 7095:143-152, 2011.
2. D. Huang, Y. Huang, Y. Bai, D. Chen, R. Hofestadt, C. Klukas, M. Chen MyBioNet: interactively visualize, edit and merge biological networks on the Web. Bioinformatics, 2011. Online: DOI: 10.1093/bioinformatics/btr557
3. M. Chen, S. Hariharaputran, R. Hofestädt, B. Kormeier, S. Spangardt Petri net models for the semi-automatic construction of large scale biological networks. Natural Computing, 10(3):1077-1097, 2011. Online: DOI: 10.1007/s11047-009-9151-y.

IN SILICO RECONSTRUCTION OF MULTI-PROTEIN COMPLEX INTERACTING WITH THE COMMON REGULATORY REGIONS IN THE HUMAN *CYP1A1/1A2* INTERGENIC SEQUENCE

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

Key words: *transcription, CYP1A1, CYP1A2, multi-protein complex, coregulatory proteins*

Motivation and Aim: CYP1A1 and CYP1A2 are involved in detoxification of drugs, nutrients, and environmental pollutants. The human *CYP1A1* and *CYP1A2* genes are located in a head-to-head orientation, sharing a 5'-flanking region which includes a common regulatory elements. Identification of regulatory proteins and signal transduction pathways involved in *CYP1A1* and *CYP1A2* transcription regulation is necessary to predict potential effects of xenobiotics and to develop optimal treatment strategies.

Methods and Algorithms: The list of transcription factors (TFs) involved in *CYP1A1* and *CYP1A2* transcription regulation was compiled manually from scientific publications. The set of coregulatory proteins, which may be involved in regulation, was formed using a novel database of transcriptional regulators - TrDB [1]. The coregulatory proteins were ranked by the number of known interactions with TFs

Results: According to published data, at least six well known TFs, including AhR/Arnt heterodimer, interact with the regulatory elements in the human *CYP1A1/1A2* intergenic sequence. More then 100 different transcriptional regulators, which may interact with these six TFs, were extracted from TrDB. Among them two coregulatory proteins (NCOA1 and PPARGC1A) are known to participate in protein-protein interactions with four of six TFs and two proteins (PPARBP and PPARGC1A) are known to interact with three of six TFs. Besides, six coregulatory proteins may interact with two of six TFs. The manual verification of the top-ranked coregulatory proteins was carried out using the PubMed database. A number of the PubMed abstracts confirming an activatory role of NCOA1, PPARGC1A, PPARBP, PPARGC1A in the human CYP1A1 or CYP1A2 transcription were found.

Conclusion: The ranked set of transcriptional regulators which may be components of multi-protein complex, controlling *CYP1A1* and *CYP1A2* expression was compiled using TrDB. Checking the biological roles of the top-ranked proteins using the PubMed database demonstrated the effectiveness of our approach to search for potential coregulators. Further experimental verification may help to identify all members of multi-protein coregulatory complex.

Acknowledgements: SB RAS project №136, RAS projects №№ 6.8, 28, 30.29., Russia's President's project №-5278.2012.4.

References:

1. Ignatieva E.V. (2012) TrDB: a database of the human, mouse, and rat transcriptional regulators and its potential applications in systems biology, *This issue.*

APPLICATION OF THE ANDVISIO COMPUTER SYSTEM TO THE INTERPRETATION OF BIOLOGICAL FUNCTIONS OF PROTEINS, DIFFERENTIALLY EXPRESSED IN BRONCHOALVEOLAR LAVAGE OF MICE AFTER A ONE-TIME INTRANASAL ADMINISTRATION OF SiO_2 NANOPARTICLES

Ignatieva E.V.*, Ivanisenko V.A., Tiys E.S., Demenkov P.S., Moshkin M.P., Peltek S.E.
Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
e-mail: eignat@bionet.nsc.ru
** Corresponding author*

Key words: *ANDVisio, associative networks, biological effects of nanoparticles, SiO_2*

Motivation and Aim: Recently acute immune response on the intranasal application of nanoparticles of SiO2 (Tarkosil 25) in mice of two strains was investigated [1]. Differentially expressed proteins were found in bronchoalveolar lavage fluid (BAL) after SiO2 nanoparticles administration. To characterize some of the molecular events, stimulated in mouse respiratory tract in response to a one-time intranasal administration of SiO_2 nanoparticles, an associative network, which represents molecular relationships between SiO_2, differentially expressed proteins, and other proteins, metabolites, and molecular processes, was reconstructed and analyzed.

Methods and Algorithms: An associative network was reconstructed with the use of the ANDVisio computer system, which includes a database of knowledge and facts extracted automatically from PubMed together with data from more than 20 public databases [2]. The ANDVisio system provides users with capabilities for data visualization and analysis.

Results: The associative networks including differentially expressed BALB/c and C57Bl/6 mice BAL proteins were reconstructed with the use of the ANDVisio system. Before the manual verification these networks contained (respectively) 338 or 102 proteins, more than 600 or 200 metabolites and 116 or 53 processes or metabolic pathways. After verification and analysis of network interactions the objects (proteins, metabolites and processes) with the largest number of links were identified: i) TNFa, FGF2, PGH2; ii) calcium atom; iii) carbohydrate metabolic process, apoptosis and the extracellular actin-scavenger system.

Conclusion: ANDVisio computer system enables to reveal functionally important associations between biomolecules. Two of differentially expressed proteins participate in clearing the actin from extracellular fluids. Several enzymes are involved in carbohydrate metabolism. The biological effects of SiO_2 nanoparticles may be mediated through TNFa, FGF2, PGH2 and calcium ions.

Acknowledgements: SB RAS projects № 57, 136, RAS projects № 6.8, 28, 30.29., Russia's President's project №-5278.2012.4.

References:

1. Moshkin M.P. et al. (2011) Acute immune response on the intranasal application of nanoparticles of SiO2 (Tarkosil 25) in mice of two strains, Nanotechnologies in Russia, **9–10**: 89-98
2. Demenkov P.S. et al. (2008) Associative network discovery (AND) - the computer system for automated reconstruction networks of associative knowledge about molecular-genetic interactions, Computational technologies, **13**: 15-19

TrDB: A DATABASE OF THE HUMAN, MOUSE, AND RAT TRANSCRIPTIONAL REGULATORS AND ITS POTENTIAL APPLICATIONS IN SYSTEMS BIOLOGY

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

Key words: *transcriptional regulators, database, applications in systems biology*

Motivation and Aim: Transcriptional regulation in eukaryotes is very complex and involves a great number of regulatory proteins. TrDB (DataBase of Transcriptional regulators) is intended to integrate information on human, mouse and rat transcription factors and coregulatory proteins, providing data for investigations in bioinformatics and systems biology.

Methods and Algorithms: Data were extracted from public resources: i) Entrez Gene Database; ii) TcoF-DB; iii) CREMOFAC; iv) an atlas of combinatorial transcriptional regulation in mouse and man [1].

Results: Currently TrDB includes data on human, mouse, and rat transcription factors and coregulators (more than 4000 entries). The definition of an entry is based on protein and gene identifiers, gene official symbol and official full name, chromosome location. Data on protein-protein interactions between transcription regulators and tissue specificity scores are also compiled. The distribution of genes encoding transcriptional regulators was analyzed in the human genome. The regions with high and low gene densities have been found. It was revealed that chromosome 19, which has the highest gene density (but not gene content) of all human chromosomes, has the highest content and density of genes, encoding transcriptional regulators. TrDB was also used for functional interpretation of ChIP-chip data. As an example the list of 1141 genes which promoters were occupied by SREBP-1 in a human hepatocyte cell line [2] was analyzed, and 102 transcriptional regulators were revealed. Genes encoding well known transcription factors with important biological functions were found among them: *TP53, SP1, SP3, SREBF2, GTF2A1, GTF2B, TAF1B, GTF2E2, TAF1.*

Conclusion: TrDB may be used for generation of new knowledge in bioinformatics and systems biology: i) functional annotation of gene lists and gene clusters, obtained from microarray, ChIP-chip or ChIP-seq experiments; ii) searching the regulatory interactions in genetic networks; iii) analysis of spatial organization of genomes.

Acknowledgements: SB RAS project №136, RAS projects №№ 6.8, 28, 30.29., Russia's President's project №-5278.2012.4.

References:

1. Ravasi T. et al. (2010) An atlas of combinatorial transcriptional regulation in mouse and man, *Cell*, **140**: 744-752.
2. Reed B.D., et al. (2008) Genome-wide occupancy of SREBP1 and its partners NFY and SP1 reveals novel functional roles and combinatorial regulation of distinct classes of genes, *PLoS Genet*, **4**: e1000133.

ANALYSIS OF SNP DISTRIBUTION AND INTER-SNP DISTANCE IN THE HUMAN GENOME

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

Key words: *SNP distribution, the human genome, 1000 Genomes Project*

Motivation and Aim: Mutation frequencies are known to vary along a nucleotide sequence. Nucleotide positions with an exceptionally high mutation frequency are called hotspots [1]. The results of the pilot phase of the 1000 Genomes Project were approximately 15 million SNPs, 55% of which were previously undescribed [2]. But the genome-scale patterns of SNP distribution have been poorly investigated yet. Study of the SNP distribution and inter-SNP distance in the human genome is of great importance for experimental researches. Sensitivity and specificity of most experimental methods of SNP typing are strongly decreased from the presence of additional SNP variants in the neighborhood with the target SNP. Using the data from the 1000 Genomes Project, we have analyzed the pattern of SNP distribution among the different regions of the human genome and the DNA context features of nucleotide sequences with adjacent SNPs.

Results: It was found that about 1.3% of SNPs occur in neighboring positions (adjacent SNPs). It is 3 times more, than expected accident frequency for two adjacent SNPs. In 0.8 % of cases SNPs are separated by one nucleotide. SNPs density was dependent on their localization across the different parts of the gene. Low SNPs density was found in the vicinity of transcription start sites, 5'- and 3'- parts of introns. This observation is in good agreement with important functional roles of these regions. General DNA context features of the stretches with adjacent SNPs were detected: the frequency of CpG was found to be higher among this two adjacent SNPs and AT-content was found to be higher in 5' and 3' flanks (3-5 bp).

Conclusion: SNPs are localized non-uniformly along the human genome. Contrary to expected stochastic SNP distribution (one SNP occurs on average every 268 bp), more than half of all SNPs (69%) occurred at distances less than 250 bp. One of the most popular tools for screening of SNP variants, high-resolution melting curve analysis (HRM) is especially vulnerable, because for SNP screening, DNA fragments of 150–250 bp are usually used [3]. The possibility of existence of additional SNP variants in melting fragment requires careful consideration for improvement of new HRM assay specificity.

Acknowledgements: SB RAS project № 136, RAS projects № 6.8, 28, 30.29., Russia's President's project №-5278.2012.4.

References:

1. Rogozin I.B. et al. (2003) Computational analysis of mutation spectra, *Brief. Bioinform.* **4**:210-27.
2. 1000 Genomes Project Consortium et al. (2010) A map of human genome variation from population-scale sequencing, *Nature,* **467**:1061-1073.
3. Vossen R.H. et al. (2009) High-resolution melting analysis (HRMA): more than just sequence variant screening, *Hum. Mutat.*, **30:**860-866.

RECONSTRUCTION OF THE ASSOCIATIVE GENETIC NETWORKS BASED ON INTEGRATION OF AUTOMATED TEXT-MINING METHODS AND PROTEIN-LIGAND INTERACTIONS PREDICTION

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

Motivation and Aim: Genetic network is a group of coordinated functioning genes. The associative genetic network is a genetic network reconstructed *in silico* with the use of formalized rules for identification for identification of interactions between molecular-genetic objects from the texts of scientific publications and databases. Manual analysis of such data by experts has high degree of accuracy but is very time-consuming therefore it makes timely the task of development of interactive tools for analysis of a full-text articles which can be used for reconstruction of the associative genetic networks. The aim of this work was a development of the integrated computer system for the reconstruction of associative genetic networks based on interactive analysis of scientific literature and prediction of protein ligand interactions.

Methods and Algorithms: Automated extraction of information about molecular-genetic, genetic-genetic, metabolite-genetic and other types of interactions in text was performed using the text-mining methods we developed. For text-mining, we also used previously developed by us thesaurus with the names of proteins, genes, metabolites, diseases, microRNAs, biological pathways, cells and organisms. A web-based interface allows text-mining module to work in a real-time mode with textual files (in a *.pdf or *.txt format) uploaded by experts and to demonstrate results of data extraction in user-friendly way as well. The integrated system for prediction of protein-ligand interactions, based on recognition of functional sites in a tertiary structure of proteins feature, allows to get new previously unknown interactions.

Results: An integrated computer system for reconstruction of associative genetic networks based on interactive analysis of scientific literature and prediction of protein-ligand interactions (ITMSys) was developed. The ITMSys software is equipped with tools for reconstruction, visualization of genetic networks, prediction of new interactions and web-based interface. The ITMSys ensures analysis of full-text articles connected with genetic networks.

Conclusion: Developed system can be used by experts for the acceleration of genetic networks reconstruction process, as well as in the other fields of science related to automated analysis of the scientific texts.

Acknowledgements: Work is supported in part by Russian Ministry of Education and Science № 07.514.11.4003 and 14.740.11.0001.

ASSOCIATIVE NETWORK DISCOVERY SYSTEM (ANDSYSTEM): AUTOMATED LITERATURE MINING TOOL FOR EXTRACTING RELATIONSHIPS BETWEEN DISEASES, PATHWAYS, PROTEINS, GENES, microRNAs AND METABOLITES

Ivanisenko V.A.*[1, 2], Demenkov P.S.[1], Ivanisenko T.V.[1], Tiys E.S.[1]
[1] *Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia;*
[2] *PBSoft Llc, Novosibirsk, Russia*
e-mail: salix@bionet.nsc.ru
** Corresponding author*

Motivation and Aim: Work with scientific literature as well as factual databases is required for research in every area of knowledge. In order to enable formulation of problems and hypotheses scientists are studying the existing pool of scientific knowledge. The size of this pool is immense and expands exponentially. Nearly 40 % of such data are biomedical in nature. For example, the Pubmed database contains over 20 million of scientific abstracts, and their number increases annually by 1 million per year. Reading of all of them, even if it will take only 3 minutes on each, would take more than 200 years. In this way development of tools for automated literature analysis (text-mining) becomes a timely task.

Methods and algorithms: The text-mining algorithms implemented in the ANDSystem are based on semantic patterns. The improved method for semantic analysis of biological texts employs a link grammar parser combined with semantic patterns was developed. Also we have developed original methods for construction of pathways and cells vocabularies. The ANDSystem is provided with methods for automated reconstruction of associative gene networks, which describe semantic relationships between molecular-genetic objects (proteins, genes, metabolites and others), biological processes, and diseases.

Results: The ANDSystem was developed for the purpose of scanning literature for extracting relationships between diseases, pathways, proteins, genes, microRNAs and metabolites. The ANDSystem incorporates utilities for automated extraction of knowledge from Pubmed published scientific texts and analysis of factographic databases. The ANDCell database contains information on molecular-genetic events retrieved from texts and databases. The ANDVisio is a user's interface to the ANDCell database stored on the remote server. It provides graphic visualization, editing and search features as well as possibilities to save an associative gene networks in different formats resulting from user's request. The ANDVisio is provided with various tools supporting filtering by object types, relationships between objects and information sources; graph layout; search of the shortest pathway; cycles in graphs.

Conclusion: The ANDSystem can assist in the interpretation of complex multifactorial experimental data. In particular, the ANDSystem was used for the analysis of proteomic experimental data. For example, with the ANDSystem was reconstructed and analyzed networks of molecular and genetic interactions of proteins of Helicobacter pylori, differentially expressed in different strains isolated from patients with chronic gastritis and gastric tumors. On the example of the establishment of interactions between human proteins and proteins of hepatitis C virus was shown a high accuracy of the method of knowledge extraction from texts.

Avaibility: The ANDSystem is available by request to developers.

Acknowledgements: Work is supported in part by 7th Framework Programme (FP7) project "New Algorithms for Host Pathogen System Biology" SYSPATHO № 260429.

SUPPRESSION OF SUBGENOMIC HCV RNA BY NS3 PROTEASE ANTIVIRALS IN CELLS: A BASIC STOCHASTIC MATHEMATICAL MODEL

Ivanisenko N.V. [1, 2], Mishchenko E.L.*[1], Akberdin I.R.[1], Demenkov P.S.[1], Likhoshvai V.A.[1, 2], Kolchanov N.A.[1, 2], Ivanisenko V.A.[1]
[1] *Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia;*
[2] *Novosibirsk State University, Novosibirsk, Russia*
e-mail: elmish@bionet.nsc.ru
** Corresponding author*

Key words: *hepatitis C virus, subgenomic replicon, mathematical modeling, drug efficacy*

Motivation and Aim: hepatitis C virus (HCV) is a severe causative agent of liver disease frequently resulting in cirrhosis and hepatocellular carcinoma. Mathematical modeling is a promising tool for the evaluation of the efficiency of new potential anti-HCV drugs on the viral genome replication. Early we have proposed a mathematical model for suppression of subgenomic HCV RNA replication in the presence of the inhibitors of various types. However, the model described the kinetics of viral RNA suppression by inhibitors during a short time, not over some days [1]. The aim of this work was to construct a new mathematical model describing the kinetics of the suppression of viral RNA by the NS3 protease inhibitors in cells during the entire time of the experimental kinetics of viral RNA.

Results: **A** stochastic mathematical model for subgenomic HCV replicon replication in Huh-7 cells in the presence of the HCV NS3 protease inhibitors was developed for the first time. A hallmark feature of the model was incorporation of the wild-type and the drug-resistant viral RNAs into the cell. The model described the experimental kinetics of the viral RNA suppression during the entire observation up to 15 days. It took into account the fact that mutant drug-resistant replicons preexist in a treatment naпve replicon population [2], also the stochastics of the replication-degradation of viral RNA.

Conclusion: A consideration of the drug-resistant forms of viral RNAs, along with random variation in RNA replication-degradation, was decisive in the achievement of an accurate description of the experimental kinetics of viral suppression by the NS3 protein inhibitors, when viral RNA number/cell was reduced to a few molecules. At high concentrations of viral RNA in the cell (the initial time of inhibitor action), consideration of drug-resistant forms of viral RNA and of the stochastic proved to be irrelevant to the description of the kinetics.

References:

1. E.L. Mishchenko et al. (2007) Mathematical model for suppression of subgenomic hepatitis C virus RNA replication in cell culture, *J.Bioinform. Comput. Biol.,* **5**: 593-609.
2. M. Robinson et al. (2011) Preexisting drug-resistance mutations reveal unique barriers to resistance for distinct antivirals, *Proc. Natl. Acad. Sci. USA.,* **108**: 10290-10295.

FEATURES OF hsa-miR-1279 BINDING SITES IN PROTEIN CODING SEQUENCE OF *PTPN12, MSH6, ZEB1* GENES

Ivashchenko A.T.*, Issabekova A.S., Berillo O.A., Khailenko V.A.
Al-Farabi Kazakh National University, Almaty, Kazakhstan
e-mail: a_ivashchenko@mail.ru
** Corresponding author*

Key words: *microRNA, mRNA, PTPN12 gene, MSH6 gene, ZEB1 gene*

Motivation and Aims: miRNA binding sites with mRNA are located within 5'UTR, CDS and 3'UTR. Near half the total number of miRNA sites is in protein coding sequence (CDS) that influences high attention to these sites. Many miRNAs are similar or not significantly different in closely related and phylogenetically distant species. We supposed presence of the similar miRNA binding sites in orthologous target genes. The aim of our investigation determines conservation of one miRNA site in CDS of orthologous genes.

Methods: Nucleotide sequences of mRNAs of genes and hsa-miR-1279 were obtained from Genbank (http://www.ncbi.nlm.nih.gov) and miRBase (http://www.mirbase.org) respectively. The free energy value (ΔG) of hybridization of hsa-miR-1279 with studied mRNAs was calculated using RNAHybrid 2.1 program. Nucleotide and amino acid sequences variability diagrams were visualized by WebLogo program.

Results: hsa-miR-1279 binding sites with mRNA are in human protein coding sequence of *PTPN12*, *MSH6*, *ZEB1* genes, which are included to tyrosine phosphatase family, DNA repair family, zink-finger family accordingly. These genes are involved in development of breast, colon, esophageal, stomach, lung, prostate, ovary cancer and *etc*. Polynucleotides of mRNAs in binding sites encode TKEQYE, EGSSDE, GEKPYE oligopeptides specific for *PTPN12*, *MSH6*, *ZEB1* genes respectively. mRNAs of orthologous genes from 20 animal species for human *PTPN12*, *MSH6*, *ZEB1* genes in miR-1279 binding site consist of homologous polynucleotides encoding TKEQYE, EGSSDE, GEKPYE oligopeptides in corresponding proteins. Actually other mRNAs of human genes from tyrosine phosphatase family, DNA repair family, zink-finger family have no has-miR-1279 sites. mRNAs of many genes in zink-finger gene family have from one up to twelve sites coding GEKPYE hexapeptide, but these sites weakly interact with hsa-miR-1279. It is connected with replacement in third position of polynucleotide triplets coding GEKPYE hexapeptide. Binding site in mRNA of *PTPN12* gene is 5'-dominant canonical site and sites in mRNAs of *MSH6*, *ZEB1* genes are 3'-compensatory sites.

Conclusion: Polynucleotides in hsa-miR-1279 binding sites in CDS mRNAs of human *PTPN12*, *MSH6*, *ZEB1* genes and in CDS mRNAs of their orthologs are from more than 20 animals encode conservative TKEQYE, EGSSDE, GEKPYE oligopeptides.

Availability: Obtained data have proved hsa-miR-1279 role in translation regulation of *PTPN12*, *MSH6*, *ZEB1* genes and may be useful in modulation of their expressions.

Acknowledgements: This study was supported by grant of Ministry of Education and Science of Kazakhstan.

CHALLENGES ON LARGE-SCALE COMPUTATIONAL PHYLOGENETICS

Izquierdo-Carrasco F.*, Stamatakis A.
IHeidelberg Institute for Theoretical Studies, Heidelberg, Germany
e-mail: Fernando.Izquierdo@h-its.org
** Corresponding author*

Key words: *computational phylogenetics, high performance computing, memory requirements, RAxML, Maximum Likelihood*

Motivation and Aim: The rapid accumulation of molecular sequence data, driven by the adoption of Next-Generation sequencing technologies, poses new challenges for large-scale maximum likelihood-based phylogenetic analyses. Some of these computational challenges are: the scalability of search algorithms, and the high memory requirements for computing the likelihood. Based on our experience with the user community of RaxML, memory-shortages (as opposed to CPU time limitations) are currently the prevalent problem regarding resource availability, that is, lack of memory hinders large-scale biological analyses.

Methods and Results: We develop a new search strategy that can reduce the time required for tree inferences by more than 50 % while yielding equally good trees (in the statistical sense) for well-chosen starting trees [1].

In order to reduce memory requirements, we introduce and implement a novel, general, and versatile concept to trade memory consumption for additional computations in the likelihood function which exhibits a surprisingly small impact on overall execution times. When trading 50 % of the required RAM for additional computations, the average execution time increase because of additional computations amounts to only 15 % [2].

Conclusion: All concepts presented here are sufficiently generic such that they can be applied to all programs that rely on the phylogenetic likelihood function, for instance RAxML-Light [3]. Thereby, the approaches we have developed will contribute to enable large-scale inferences of whole-genome phylogenies.

References:

1. Izquierdo-Carrasco F., Smith S.A. and Stamatakis A.. (2011) Algorithms, Data Structures, and Numerics for Likelihood-based Phylogenetic Inference of Huge Trees, *BMC Bioinformatics,* **12(1)**: 470+.
2. Izquierdo-Carrasco F. et al. (2012) Trading memory for running time in phylogenetic likelihood computations, *In Proceedings of Bioinformatics 20012*.
3. A. Stamatakis, A.J. Aberer, C. Goll,S.A. Smith, S.A. Berger, F. Izquierdo-Carrasco (March 2012): "RAxML-Light: A Tool for computing TeraByte Phylogenies", Heidelberg Institute for Theoretical Studies, *Exelixis-RRDR-2012-3*.

IN SILICO EVIDENCE OF THE NOTCH SIGNALLING PLAYERS IN LEUKEMIA

Jamil K., Jayaraman A., Sabeena K.M. Kakarala, Khan M.
Centre for Biotechnology and Bioinformatics, School of Life sciences,
Jawaharlal Nehru Institute of Advanced Studies (JNIAS)
Budha Bhawan, 6th M.G. Road , Secunderabad, 500003, A.P. India
Correspondence: Kaiser.jamil@gmail.com

Abstract: One of the most crucial signaling pathways involved in hematopoieses, especially in T-lineage development, is Notch signaling. Mutations in the PEST and HD domains of Notch1 have been associated with increased ligand independent Notch activity. Notch signaling plays a critical role in cell fate determination and maintenance of progenitors in many developmental systems. Our aim was to understand the protein interaction network which was done using STRING software, using Notch1 as query we generated a model to assess the significance of Notch1 associated proteins in Acute Lymphoblastic Leukemia (ALL).

Construction of Protein-Protein Interaction Network: To infer the interactions of NOTCH1 with other proteins, we used String database v9. STRING (Search Tool for the Retrieval of Interacting Genes, available at : http://string-db.org/ Szklarczyk et al., 2011) is a database of functional associations that have been pre-computed and derived from a wide range of sources such as high-throughput experimental data, literature and database mining, analyses of co-expressed genes and computational predictions. The database interactions are based on a scoring framework and the output interactions having a single confidence score per prediction.

We further analyzed the expression levels of the genes encoding hub proteins, using Oncomine database, to determine their significance in leukemogenesis. Of the forty two hub genes, sixteen were underexpressed and eleven genes were overexpressed in T-cell ALL in comparison to their expression levels in normal cells. Using a few signature genes in our study one may provide new insights into the abnormal hematopoietic process as these genes are involved in Notch signaling and cell adhesion processes. It is evident that experimental validation of the protein interactors in this study in leukemic cells could help in the identification of new diagnostic markers for leukemia and may also be useful in development of effective therapeutic measures.

MODELLING GENE REGULATION OF MORPHOGENESIS IN THE SEA ANEMONE *NEMATOSTELLA VECTENSIS*

Kaandorp J.A.
Section Computational Science University of Amsterdam
e-mail: J.A.Kaandorp@uva.nl
URL: http://www.science.uva.nl/~jaapk/

In this presentation we couple a model of gene regulation to a cell-based model of embryogenesis. In this case study we are collecting recently published spatio-temporal and quantitative gene expression patterns from various developmental stages in Nematostella in a spatial data base (``the virtual embryo''). We use this three-dimensional data for constructing a mathematical model of the regulatory network and for inferring regulatory network parameters. The regulatory network is modelled using a set of coupled reaction-diffusion equations, while the model parameters are inferred from the data base using optimization techniques. The regulatory network model is coupled to a biomechanical model of cell movement.

A COMPUTER SYSTEM FOR KINETIC ANALYSIS OF GENE NETWORKS

Kabakov M.A.*[1, 2], Timonov V.S.[1, 3], Gunbin K.V. [2]
[1] *Novosibirsk State University, Novosibirsk, Russia;*
[2] *Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia;*
[3] *Siberian State University of Telecommunications and Information Sciences, Novosibirsk, Russia*
e-mail: pine2@mail.ru
** Corresponding author*

Key words: *metabolic pathway, rate-limiting steps, enzyme kinetics*

Motivation and Aim: It is well known that the rate determining (limiting) steps in chemical and biochemical processes are the slowest steps. For a biological point of view, it is of importance to know which steps in the processes, such as a metabolic pathway or signaling cascade, are the rate determining steps, because these steps could be the primary targets for natural selection. Thus, the strategic aim is to construct and test the integrative software for analysis of the connectivity between biochemical process kinetic and evolutionary modes of enzymes. Recently we implement and test the web-available pipeline for analysis of evolutionary modes of proteins and genes [1]. Now we implement the software for automatic/manually-assisted analysis of biochemical processes kinetic in order to determine all possible rate-determining steps in the processes.

Methods and Algorithms: Metabolic pathways can be represented by a directed graph, where edges are reactions catalyzed by enzymes, and vertices are substrates and products of reactions. In this work, the graph topologies were taken from KEGG database [2], the kinetic data were taken from Sabio-RK [3] and BRENDA [4] databases. The Java-APIs of these three databases were used for information retrieval. The turnover numbers and/or the enzyme activity were chosen as comparable measures of enzyme kinetics (CMEK). Finding CMEK values for various enzymes is a nontrivial task, because only limited number of biochemical reactions is well studied. We used the CMEK of enzymes for closely related organisms (retrieved by BLAST) if the CMEK information of analyzed enzyme is unavailable. In cases when the CMEK information cannot be found the program displays the caution and requests the data for missing CMEK data.

Results: Using above mentioned methodology a Java-program for analyzing the metabolic pathways kinetics was made. We tested this program on tricarboxylic acid cycle and obtained the well known data about the importance of citrate synthase, α-ketoglutarate dehydrogenase, succinate dehydrogenase and isocitrate dehydrogenase catalyzing reactions.

Conclusion: A Java-program for analyzing the metabolic pathways kinetics was made.

Availability: Java program available upon request

Acknowledgements: The work supported by SB RAS project 136; RAS project 6.8.

References:

1. Gunbin K.V. et al. (2010) A computer system for the analysis of molecular evolution modes of protein-encoding genes (SAMEM), *Moscow University Biological Sciences Bulletin*, **65**:142-144.
2. Kanehisa M. et al. (2012) KEGG for integration and interpretation of large-scale molecular datasets, *Nucleic Acids Res.*, **40:** D109-D114.
3. Wittig U. et al. (2012) SABIO-RK--database for biochemical reaction kinetics, *Nucleic Acids Res.,* **40:** D790-D796.
4. Scheer M. et al. (2011) BRENDA, the enzyme information system in 2011, *Nucleic Acids Res.*, **39:** D670-D676.

SOMATIC COPY-NUMBER ALTERATION CAN HELP PREDICT THE TISSUE ORIGIN OF CANCERS OF UNKNOWN PRIMARY

Kaczkowski B.*[1], Sinha R.[2], Schultz N.[2], Sander C.[2], Nielsen F.C.[3], Winther O.*[4]
[1] The Bioinformatics Centre, Department of Biology and Biomedical Research and Innovation Centre, Copenhagen University, Ole Maaloes Vej 5, 2200 Copenhagen, Denmark;
[2] Computational Biology Center, Memorial Sloan-Kettering Cancer Center, New York, New York 10065, USA;
[3] Department of Clinical Biochemistry, Copenhagen University Hospital, Blegdamsvej 5, 2100 Copenhagen, Denmark;
[4] DTU Informatics, Technical University of Denmark, 2800 Lyngby, Denmark
e-mail: bok@binf.ku.dk
** Corresponding author*

Key words: *cancer of unknown primary, copy number alteration, classification*

Motivation and Aim: Cancer of Unknown Primary (CUP) is a highly aggressive, heterogeneous disease that represents 3-5% of all new cancer cases. Additionally, not knowing the origin of the cancer poses a challenge for treatment of CUP patients. Here, we aim to build a classifier based on DNA copy number profiles that can aid the prediction of primary site of CUPs.

Methods and Algorithms: We used the DNA copy number profiles from 3573 tumours of 19 origins, 3796 normal tissue samples and 639 cancer cell lines. The pre-processed, segmented DNA copy number data were transformed into matrix format by dividing the genome into 1,030 chromosomal regions and assigning the copy number value for each of them. Subsequently we trained two classifiers based on linear discriminant analysis (LDA) and k-nearest neighbour (KNN). The classifiers were trained on primary tumour data and the performance was estimated using cross validation. We define cases where LDA and KNN agreed as high confidence predictions.

Results: The classifier could predict the origins of 72% primary tumours with 95% accuracy. As expected, the classifier was not able to predict the origin of the tumours with low number of somatic copy-number alterations; those were appropriately predicted as normal tissue. We applied the classifier to predict the origin of the cell lines and 33% of cancer cell lines were predicted with high confidence. Glioma, kidney, head & neck, breast and colon cancer cell lines were predicted with very high accuracy, whereas lung and pancreas cancer cell lines were mostly misclassified.

Conclusion: We propose, that DNA copy number profiles can be used to predict the primary site of CUP and can complement available messengerRNA and miRNA based classifiers.

Availability: The copy number data can be downloaded from www.broadinstitute.org/ccle/ and tcga-data.nci.nih.gov/tcga/. The R code to reproduce the results can be obtained by request at bok@binf.ku.dk.

NETWORK INTERPRETATION AND META-ANALYSIS OF INDEPENDENT COMPONENTS EXTRACTED FROM BREAST CANCER TRANSCRIPTOMES

Kairov U.Ye.*[1, 2], Zinovyev A.Yu.[3], Karpenyuk T.A.[1], Ramanculov Ye.M.[2]
[1] *Kazakh National University after Al-Farabi, Almaty, Kazakhstan;*
[2] *National Center for Biotechnology of the Republic of Kazakhstan, Astana, Kazakhstan;*
[3] *Institute Curie, Paris, France*
e-mail: andrei.zinovyev@curie.fr
** Corresponding author*

Key words: *Independent Component Analysis, microarrays, transcriptome, gene network*

Motivation and Aim: The high-throughput genomic technologies and particularly the microarray technology have a major impact on studying cancer. Huge amount of microarray data requires application of reproducible statistical approaches. In our study we aimed to apply Independent Component Analysis (ICA) [1] to do meta-analysis of breast cancer gene expression data and extract meaningful molecular signals in the form of gene networks.

Methods and Algorithms: We used raw microarray data (*.CEL files) of four different breast cancer datasets GSE1456, GSE2034, GSE2990, GSE3494 from the Gene Expression Omnibus database [2]. The microarrays were normalized by GCRMA and processed using R 2.8.1 software [3] and Matlab2009b [4]. Matlab version of Icasso package [5] with ICA algorithm implementation was used to analysis of independent components. Construction and visualization of gene networks and graphs was performed using the Cytoscape [6], BiNoM plug-in [7] and HPRD database [8].

Results: We identified from 7 to 8 reproducible components in all four breast cancer datasets. We developed graph-based approach to meta-analysis and interpretation of these independent components such that each of them was associated with a small gene network. Using analysis of these networks, we provided a tentative interpretation of stably reproducible components. Thus, we found that various factors such as proliferation, immune response, contamination of tumor cells by lymphocytes and normal tissues affect gene expression in breast cancer.

References:

1. P.Comon. (1994): Independent Component Analysis: a new concept?, Signal Processing, 36(3):287–314.
2. http://www.ncbi.nlm.nih.gov/geo/
3. http://www.bioconductor.org/
4. http://www.mathworks.com/
5. J.Himberg, A.Hyvarinen and F.Esposito. (2004): Validating the independent components of neuroimaging time series via clustering and visualization., Neuroimage, 22(3):1214-1222.
6. M.Cline, M.Smoot, E.Cerami et.al. (2007): Integration of biological networks and gene expression data using Cytoscape., Nature Protocols, 2, 2366 - 2382.
7. A.Zinovyev, E.Viara, L.Calzone, E.Barillot. (2008): BiNoM: a Cytoscape plugin for manipulating and analyzing biological networks., Bioinformatics, 24(6):876-877.
8. Prasad, T. S. K. et al. (2009): Human Protein Reference Database - 2009 Update. Nucleic Acids Research. 37, D767-72.

PROTEIN FOLDING TURBULENCE

Kalgin I.V.[1], Chekmarev S.F.*[1, 2]
[1] *Novosibirsk State University, Novosibirsk, Russia*
[2] *Kutateladze Institute of Thermophysics, SB RAS, Novosibirsk, Russia*
e-mail: chekmarev@itp.nsc.ru
** Corresponding author*

Key words: *molecular dynamics, protein folding flows, turbulent flows, eddies, self-similarity, structure functions, cascades of the structural transformations*

Motivation and Aim: Protein folding and hydrodynamic turbulence are two long-standing challenges, in molecular biophysics and fluid dynamics, respectively. The theories of these phenomena have been developed independently and used different formalisms [1,2]. However, as has recently been observed [3,4], folding flows of a protein can also be filled with vortices, similar to turbulent flows of a fluid. Here, we characterize the folding flows in terms accepted in hydrodynamic turbulence [5] and examine how far the similarity between the protein folding flows and turbulent flows of a fluid extends.

Methods and Algorithms: Using molecular dynamics methods, two model proteins are studied: a fyn SH3 domain (C-alpha model) and beta3s miniprotein (all-atom model). Folding fluxes are calculated in a reduced (3D) space of orthogonal collective variables, which is characterized either by the numbers of native contacts between protein sections (a SH3 domain) or hydrogen bonds distances (beta3s).

Results: We have found that at and below the glass transition (melting) temperature, the folding flows are surprisingly similar to turbulent flows of a liquid. The flows have fractal nature and are filled with 3D eddies. The eddies contain strange attractors, at which the tracer flow paths behave as saddle trajectories. Two regions of the space increment exist, in which the folding flux variations are self-similar with the space correlation (structure) functions being in close agreement with those in the Kolmogorov theory of turbulence [2]. In one region, the cascade of protein structural transformation is directed from larger to smaller scales (net folding), and in the other, it is oppositely directed (net unfolding).

Conclusion: The protein folding turbulence has many properties of hydrodynamic turbulence. The cascade mechanism of protein transformations plays a key role for this similarity, unifying this new phenomenon with the wave, superfluid and market turbulences.

References:

1. K.A. Dill, S.B. Ozkan, M.S. Shell, T.R. Weikl (2008) The protein folding problem, Annu. Rev. Biophys., 37: 289-316.
2. L.D. Landau, E.M. Lifshitz (1987) Fluid Mechanics (Pergamon, New York).
3. S.F.Chekmarev, A.Yu.Palyanov, M. Karplus (2008) Hydrodynamic Description of Protein Folding, Phys. Rev. Lett, 100: 018107.
4. I.V.Kalgin, M. Karplus, S.F.Chekmarev (2009) Folding of a SH3 Domain: Standard and “Hydrodynamic” Analyses, J. Phys. Chem. B, 113: 12759-12772.
5. I.V.Kalgin, S.F.Chekmarev (2011) Turbulent Phenomena in Protein Folding, Phys. Rev. E, 83: 011920.

THE ROLE OF CASEIN KINASES 1 IN PLANT CYTOSKELETON REGULATION

Karpov P.A.*, Raevsky A.V., Sheremet Ya.A., Blume Ya.B.
Institute of Food Biotechnology and Genomics NAS of Ukraine, Kyiv
e-mail: karpov.p.a@gmail.com, karpov_pavel@univ.kiev.ua
** Corresponding author*

Key words: *Casein kinases 1, D4476, docking, MD, isoforms, D4476, bioinformatics*

Motivation and Aim: Casein kinases 1 (CK1) are ubiquitously expressed in eukaryotic organisms and yeast. Animal CK1 isoforms (α, β, γ1-3, δ, and ε) from different organisms demonstrate a high level of conservativeness in catalytic domains. Their phosphate-binding site is a target for ATP-competitive inhibitors. Phosphate binding regions of animal CK1 showed their isoform complete identity. The functions and structures of plant CK1-like kinase isoforms in plant cell, and regulation of cytoskeleton are still unclear.

Methods and Algorithms: Plant homologs were identified based on blastp-scanning of the UniProt database. NJ-clustering was performed in MEGA5. 3D-models were optimized with amber3 ff and subjected MD in Amber99 ff. Molecular docking was performed in CCDC GOLD. Spatial structure analysis was performed in PyMol. Experiments were performed on GFP-labeled microtubules (MT) in cells of *Arabidopsis thaliana*

Results and Discussion: blastp-scanning of the UniProt database against catalytic domains of human and animal CK1δ revealed presence of 34 CK1 homologous in *A. thaliana*. Comparing of gene loci (based on *Tair* data) we confirmed the existence of only 18 CK1-like kinases, and uniqueness of their catalytic domains was confirmed cladistically. According to the BLAST protocol, the maximum "Score" (456) belongs to *A. thaliana* KC1D (UniProt: P42158, identity = 78 %, similarity = 92 %).

We have performed a reconstruction of the spatial structure of all catalytic domains from CK1s from *R. norvegicus* and its homologs from *A. thaliana*. After optimization in Amber3 force field and molecular dynamics (10 ns in Amber99 force field), the models were structurally superimposed. The high similarity between folding of CK1s from *R. norvegicus* and 13 (out of 18) CK1-like kinases from *A. thaliana* was confirmed. Recently, it was shown that D4476 inhibit CK1δ at 10 μM concentrations by more than 90 % and had almost no effect on the other protein kinases [1]. In our experiments on *A. thaliana* we have observed strong effect of D4476 on primary root growth and MT organization. In view of the similarity of spatial structures and amino acid compositions of catalytic domains, we specify casein kinases KC1D (At4g26100.1) and CKL2 (At1g72710.1) as the most likely targets of D4476. We have identified that plant homolog CKL6 contain 382-VSEKGRNTSRYG-394 motive in C-end region and has homology to mammalian MT-associated protein Eml4. These data confirm G. Ben-Nissan et al. experimental data [2]. It is known, that mammalian CK1δ binds and phosphorylates EB1 (MAP1) [3].

Conclusion: It seems like that in *A.thaliana* D4476-acts as inhibitor on KC1D, CKL6 and possibly CKL2 isoforms. Therefore the effects on MT organization in *A.thaliana* are likely associated with complex inhibition of mentioned CK1 isoforms. Thus, CK1s play important role in the plant cytoskeleton regulation.

References:

1. G. Rena, J. Bain, M. Elliot, P. Cohen (2004) *EMBO Rep*., 5(1): 60–65.
2. G. Ben-Nissan, W. Cui, D.-J. Kim, et al. (2003) *Plant Physiol.,* 148: 1897–1907.
3. D. Zyss, H. Ebrahimi, F. Gergely. (2011) *J. Cell Biol.*, 195(5): 781–797.

DE NOVO SEQUENCING, ASSEMBLY AND CHARACTERIZATION OF TRANSCRIPTOME IN TETRAPLOID PLANT *CAPSELLA BURSA-PASTORIS*

Kasianov A.S.*[1, 2], Logacheva M.D.[3], Oparina N.Y.[1], Penin A.A.[3]
[1] *Engelhardt Institute of Molecular Biology RAS, Moscow, Russia;*
[2] *Vavilov Institute of General Genetics, RAS Moscow, Russia;*
[3] *Lomonosov Moscow State University, Moscow, Russia*
* *Corresponding author*

Key words: *plants, polyploidy, transcriptome, sequence assembly, Capsella bursa-pastoris*

Motivation and Aim: Transcriptome sequencing data is an essential component of modern genetics, genomics and evolutionary biology. Huge improvements of sequencing technologies allowed characterization of transcriptomes in many non-model species. However, de novo assembly of transcriptomes of flowering plants is still a challenging task due to the fact that many of them are recent polyploids and thus multiple paralogs very similar to each other are present in their genomes and this hampers the assembly.

We performed sequencing and analysis of transcriptome of *Capsella bursa-pastoris*. This plant is a tetraploid with uncertain origin, being a recent allotetraploid or more ancient autotetraploid. Its close relationship with model plant species, *Arabidopsis thaliana* (*Capsella* belongs to the same family) makes *C. bursa-pastoris* a perfect model for the studies of gene and genome evolution after genome duplication events.

Methods and Algorithms: cDNA corresponding to the genes expressed *C.bursa-pastoris* flowers and inflorescences was sequenced using Illumina and 454 sequencing platforms. Also, additional set of cDNA libraries corresponding to the genes expressed in various stress conditions (cold, over-illumination, wounding) was sequenced. In total, nearly 60 millions of reads were generated. Different programs were tested (MIRA, Velvet, CLC Genomics workbench) for assembly, but none of them demonstrated the capacity to assemble paralogous genes separately because of their high similarity. We developed the algorithm for partitioning of reads into subsets corresponding to each of the paralogs. After partitioning the subsets were assembled separately thus allowing to generate separate sequences for each paralog. Sequences were annotated using BLAST2Go; after annotation each pair of paralogs was analyzed in terms of sequence divergence, the presence of intact ORF and codon usage.

Results and conclusion: Transcriptome of *C. bursa-pastoris* was sequenced and assembled using newly developed algorithm for separation of reads into subsets corresponding to the paralogous genes. Patterns of molecular evolution in paralogous genes were inferred.

HIGH PERFORMANCE COMPUTING WITH MGSMODELLER

Kazantsev F.V. *[1], Akberdin I.R.[1], Mironova V.V. [1], Podkolodnyy N.L.[1], 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: The rapidly growing data volumes for analysis in system biology demands high performance computing. This is one of the topical problem for mathematical modeling of complex molecular-genetic systems (MGS). We have represented a new version of the MGSmodeller [1] software that designed for support of MGS model simulations and analysis.

Results & Conclusions: The MGSmodeller includes modules for mathematical model reconstruction and specification routine simulation as well as for solution of inverse problems. Model reconstruction and simulation experiment specification are describing in terms of the model specification language SiBML. The complex mathematical models in the SiBML [2] is depicted by the set of MGS elementary subsystems. MGS elementary subsystem is defined by a set of biological entities and a interaction law between them. The model specification language allows to present a full information about the biological entities such as name, type (gene, RNA, protein *etc.*), localization in the compartment structure (organelles, cells, tissue *etc.*) and/or other object properties. Thus, the language allows to take into consideration a hierarchical structure and complex organization of modeling objects. On this basis we can model such properties of the MGS as mutual positions of genes, promoters and others genetical elements on DNA, anisotropy in spatial compartments, transport processes and so on. The modeling cycle with MGSmodeller on a high performance cluster has the following stages: 1) Implementation of the MGS elementary subsystems as model library that will be complex model's building blocks; 2) Specification of the complex model structure that involves compartment structure, genetical properties, associations of elementary subsystems from the library with the complex model structure elements; 3) The complex model assembling and compilation on the basis of the previous step specification (by parallel method or not); 4) Specification of the simulation experiment's protocol that may contain methods for varying of the parameters and initial conditions; 5) Calculation on the high performance cluster and accumulation results for further analysis; 6) Analysis of the results with filters, arrangements and visualization modules that give 2D, 3D graphics and/or animation of the solution. To test the MGSmodeller efficiency on the high performance cluster we used the mathematical model of auxin distribution in the root meristem where three auxin transporters are considered - PIN1-PIN3. The model represents 2D cell layout (4x20 cells) [1] and contains 240 variables and 38 parameters. The model compilation and assembling that demanded 3 hours of process time were done with the parallel method on the cluster for 15 minutes in average. A few thousands of simulations with varying parameters were done on the cluster for a few days with the single simulation time dispersion from 20 minutes till couple of hours.

References:

1. Kazantsev F.V. *et al*. *Proc. of the 6th International Conference on BGRS*, p.113.
2. Likhoshvai V.A. *et al.* (2001). *Mol. Biol., 35:1072-1079.*
3. 3. Mironova V.V. *et al*. (2012) *Annals of botany*. doi:10.1093/aob/MCS069

DIVERSION OF GENOME LOCI AND CO-LOCALIZATION PATTERNS STUDY OF THE PROTEIN FAMILIES FROM DIFFERENT FUNCTIONAL CLASSES OF THE BACTERIAL CARBOHYDRATE METABOLISM

Kaznadzey A.D.*, Shelyakin P.V.
Institute for Information Transmission Problems RAS, Moscow, Russia
e-mail: vzmisha4@gmail.com
** Corresponding author*

Key words: *carbohydrate metabolism; bacterial gene diversion; bacterial genome evolution*

Motivation and Aim: The aim of this study is to explore genome loci of the carbohydrate metabolism in bacteria. Such loci consist of genes encoding proteins which participate in biochemical transformations of carbohydrates, such as phosphorylation, hydrolysis, etc., and also in the transport and regulation of transcription. Co-localization of proteins belonging to different isofunctional families and sub-families allows us to obtain information about evolutionary compatible combinations and to assess functional compatibility for various proteins.

Methods and Algorithms: The first step was to select an appropriate classification of proteins from the bacterial carbohydrate metabolism. We considered combinations of different classification schemes, the most accurate classification system was obtained using both COG and Pfam protein identities. About 270 carbohydrate metabolism-related COG-families and about 170 Pfam-families were analyzed. Also, each protein is assigned to a large iso-functional family, such as kinases, mutases, hydrolases, etc. For the loci research we have analyzed all the co-localization cases of carbohydrate metabolism genes and built a database consisting of the "real-existing" loci, which vary in size from 2 to about 20 genes long.

For the family pairs compatibility research we found out the co-occurrence of each subfamily pair within large isofunctional family pairs.

Results: We have studied the "strength" of each loci and "loci particles" of all sizes, by sorting them based on their occurrence level, thus finding out the strongest kinds of gene unions. Our database presents such strong unions using either sub-family or family classification.

Database with 3D-graphs of subfamily co-occurence was also built. We used several types of visualization, based on different kinds of subfamily clustering. Each co-occurrence matrix was analyzed: peak criteria was suggested, a total number of peaks was calculated as well as their percentage, etc. Significance and origin specifics of each peak was estimated based on the chi-square matrix; the results were both viewed from the family and subfamily "point of view".

Acknowledgements: This is joint work with M.Gelfand.

References:

1. R.D. Finn, et al, (2010) The Pfam protein families database. *Nucleic Acids Research, Database Issue* **38**:D211-222
2. Tatusov RL, et al, (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.* **29(1)**: 22-28.

DE-NOVO DISCOVERY OF DIFFERENTIALLY ABUNDANT DNA BINDING SITES INCLUDING THEIR POSITIONAL PREFERENCE

Keilwagen J., Grau J., Paponov I.A., Posch S., Strickert M., Grosse I.*
Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Gatersleben, Germany;
Institute of Computer Science, Martin Luther University Halle–Wittenberg, Halle, Germany;
Institute of Biology II, Albert–Ludwigs–University Freiburg, Freiburg, Germany
e-mail: grosse@informatik.uni-halle.de

** Corresponding author*

Key words: *de-novo motif discovery, discriminative learning, positional preference, auxin responsive element, DREAM challenge*

Motivation and Aim: The identification of DNA binding sites has been a challenge since the early days of computational biology, and its importance has been increasing with the development of new experimental techniques and the ensuing flood of large-scale genomics and epigenomics data yielding approximate regions of binding. Many binding sites have a pronounced positional preference in their target regions, which makes them hard to find as this preference is typically unknown, and many of them are weak and cannot be found from target regions alone but only by comparison with carefully selected control sets. Several de-novo motif discovery programs have been developed that can either learn positional preferences from target regions or differentially abundant motifs in target versus control regions, but the combination of both ideas has been neglected.

Results and Conclusions: Here, we introduce Dispom, a de-novo motif discovery program for learning differentially abundant motifs and their positional preferences simultaneously. Dispom outperforms existing programs based on benchmark data and succeeded in detecting a novel auxin-responsive element (ARE) substantially more auxin-specific than the canonical ARE. Since its publication, we have endowed Dispom with more complex motif models and extended it to handle weighted input data such as ChIP-seq or BS-seq data. We have been applying Dispom to in-house and publicly available data of different transcription factors and insulators in yeasts, plants, and mammals as well as to protein-binding microarrays, where it turned out to be one of the top-scoring approaches in the corresponding DREAM challenge.

Availability: Dispom is freely available as a component of the open-source Java framework Jstacs and as a stand-alone application at http://www.jstacs.de/index.php/Dispom.

MODELING AND ANALYSIS OF DYNAMICS OF THE RIBOPYRIMIDINES *DE NOVO* BIOSYNTHESIS IN *E. COLI*

Khlebodarova T.M. [1], Akberdin I.R.*[1], Fadeev S.I. [2, 3], Likhoshvai V.A. [1, 3]
[1] *Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia;*
[2] *Sobolev Institute of Mathematics SB RAS, Novosibirsk, Russia;*
[3] *Novosibirsk State University, Novosibirsk, Russia*
e-mail: akberdin@bionet.nsc.ru
** Corresponding author*

Key words: *ribopyrimidines, mathematical model, E. coli, carbamoyl phosphate, regulation*

Motivation and Aim: Pyrimidine nucleotides as an obligatory components of RNA and DNA are essential for metabolic support of replication and transcription processes. The ribopyrimidines *de nono* biosynthesis in *E. coli* contains 19 enzymatic reactions the final products of which are ribo- and deoxyribonucleotides UTP, CTP, dCTP and TTP. The activity of seven enzymes regulating different steps of the pathway is exposed to positive and negative regulation on the feedback mechanism by biosynthesis products. It was shown in theoretical studies that models of some genetical and biochemical systems regulated by the feedback mechanism demonstrate a complex dynamical behavior including oscillation and chaos [1–3].

Methods and Algorithms: Models of enzymatic reactions were developed on basis of the mass-action law, Michaelis–Menten equation, King–Altman method and generalized Hill functions [4]. The complex model of ribopyrimidines biosynthesis was generated using web-available computer system MGSmodelsDB developed in ICG SB RAS. Values of some model parameters were extracted from published data, values of another part were estimated using manual fitting as well as the method of gradient descent. As a functional minimization the Euclidean distance of the theoretical calculations from experimental was taken. Analysis of the individual contribution of each negative feedback in the mode of ribopyrimidines biosynthesis functioning carried out by the MGSmodelsDB and STEP+ [5].

Results and Conclusion: The model of the ribopyrimidines biosynthesis taking into account all known regulatory feedbacks was developed. The numerical analysis of the model and its 254 reduced variants containing all possible combinations of regulatory interactions is performed. It was revealed that fluctuations in concentrations of the metabolites arise only in the model's variants (128 variants) in which there are first regulatory feedback controlling carbamoyl phosphate biosynthesis on the mechanism of negative feedback through the UTP. All models which don't contain first regulatory feedback demonstrated stable steady state. Cyclic mode of the ribopyrimidines biosynthesis may be due to the necessity of the conjugation between transcription and replication processes with cell division.

Availability: http://modelsgroup.bionet.nsc.ru/MGSmodelsDB/

Acknowledgements: The study was supported in part by RFBR [10-01-00717-a], grant "SYSPATHO" The study was supported in part by RFBR [10-01-00717-a], grant "SYSPATHO" [260429] of the FP7th. The study was also partially supported by the № 5278.2012.4 grant and Programs of the Presidium RAS (6.8), (30.29), the integration project of the Presidium SB RAS № 80 and the Russian Ministry of Education and Science (contracts No. P857)

References:

1. Decroly O., Goldbeter A. (1982) *PNAS,* **79**(22): 6917-21.
2. Pigolotti S., Krishna S., Jensen M.H. (2007) *PNAS,* **104**: 6533-6537.
3. Xiao M., Cao J. (2008) *Math Biosci.*, **215**: 55-63.
4. Likhoshvai V.A., Ratushny A.V. (2007) *J. Bioinform. Comput. Biol.,* **5**: 521-531.
5. Fadeev S.I. *et al*.(2006) *Proc. of the 5th International Conference on BGRS*. **2**: 118-120.

ASSOCIATIONS BETWEEN PROMOTER POLYMORPHISMS IN KEY GENES OF LIPID METABOLISM AND MIOGENESIS AND ECONOMICALLY VALUABLE TRAITS IN PIGS

Khlopova N.S.*[1], Glazko T.T.[1], Guiatti D.[2], Stefanon B.[2]
[1] *Russian State Agrarian University – MTAA named after K.A. Timiryazev, Moscow, Russia;*
[2] *University of Udine, Italy*
e-mail: khlopova.natalia@gmail.com
** Corresponding author*

Key words: *single nucleotide polymorphisms (SNP), key genes, promoter, genotyping, genotypes, association study, pigs*

Motivation and Aim: Huge sets of gene expression profiles (GEP) data, in particular for farm animals, collected over the last two decades. The main aim of such data-sets is to search for gene expression differences associated with economically valuable traits. In an earlier article we showed subdivision of GEP into constitutive and variable part according to the individual changes in gene expression. Interesting, that difference in expression of genes from the second group can be possibly explained by effect of paratypic component [1]. In a present research we checked a possible influence of genetic component on association between regulation of gene expression and development of economically valuable traits. It was done by association analysis of mononuclear promoter polymorphisms in key genes of lipid metabolism, control of myogenesis and meat and fat characteristics of two different industrial pig crosses.

Methods and Algorithms: Polymorphic promoter regions of Acyl-CoA desaturase (SCD) [-233 T/C], Leptin (LEP) [-11606 C/T; -12109 T/A], Myoblast determination protein 1 (MYOD) [-38 G/A], Myogenic factor 6 (MYF6) [-375 T/C], Osteopontin (OPN) [-24 G/A] were genotyped with KASPar system directly from genomic DNA of two- and three-breed pig crosses of Italian selection. Phenol–chloroform extraction was used to isolate DNA from muscle tissue. Statistical analysis was carried out using the SPSS 17.0 software (SPSS statistics; 2008) Statistica 7.0 (StatSoft Inc, 2004), MedCalc Software 12.0.3, Microsoft Excel.

Results: The frequencies of alleles and genotypes showed noticeable differences for all studied SNPs. Among all genes except Lep two-breed cross differ from three-breed cross with higher level of heterozygosity. Most of associations ($P<0,05$) were observed for promoter SNP of Opn gene and back fat, characteristics of meat production of two-breed cross. Interesting, that revealed associations of production traits and SNP genotypes in promoter region are breed and sex-dependent. Associations between all SNP genotypes and average daily gain in different periods of ontogenesis were observed only for three-breed cross pigs ($P < 0,05$,$P < 0,01$,$P < 0,001$).

Conclusion: Received data demonstrates that associations between genotypes in promoter region of key genes of lipid metabolism, genes controlling myogenesis and production traits among analyzed groups of pigs depend on sex, genetic background (two- and three-breed crosses) and stages of development. Obviously complicated character of genetic component influencing regulation of the key genes of economically valuable production traits can endow to variable part of GEP.

References:

1. *Glazko T. et al.,* Gene Expression Profiles in Porcine Tissues of Liver and Kidney//Journal of Life Sciences. -2011. –V.5. –P. 192-200.

TOWARDS A PUBLIC REPOSITORY FOR SYSTEMS MICROSCOPY DATA

Kirsanova C.[1], Rustici G.*[1], Neumann B.[2], Heriche J.-K.[2], Huber W.[2], Ellenberg J.[2], Brazma A.[1]

[1] *EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, CB10 1SD, UK;*

[2] *EMBL Heidelberg, MeyerhofstraЯe 1, 69117 Heidelberg, Germany*

e-mail: gabry@ebi.ac.uk

** Corresponding author*

Key words: *systems microscopy, image data, public repository*

Motivation and Aim: Systems microscopy is a new emerging field that applies multiparametric statistical analyses and mathematical modeling to imaging-derived data in order to interrogate biological processes. As with previous 'omics' technologies, systems microscopy aims at integrating data and knowledge from independent studies into a comprehensive understanding of cellular systems. Key to this goal is the development of an infrastructure that facilitates efficient generation, processing and storage of systems microscopy data. This is the scope of the Systems Microscopy Network of Excellence (http://systemsmicroscopy.eu/), a life science project spearheading a key enabling methodology based on live cell imaging for the development of next-generation systems biology. Within this project, we are developing a prototype for a public data repository, which will provide access to systems microscopy data for the broader research community and facilitate the development of analytical methods for this field.

Methods and Algorithms: We are developing a prototype database and a web interface for the storage and visualization of data generated from high-throughput discrete perturbation screens (mainly RNAi), aimed at identifying key molecular actors and their function impact within complex cellular processes.

Results: The current prototype is a human, gene-centered, non-relational database, cross-referencing various annotation databases (i.e. Ensembl). The current interface prototype supports three basic type of queries: (i) for a gene, by name or attribute, across studies; (ii) for a reagent or siRNA, by manufacturer or screen internal identifier; and (iii) for a phenotype, or multiple phenotypes, within a given study.

Conclusion: This new repository will provide easy access to systems microscopy data and integrate independent studies, adding significant value to the hard-earned primary data.

Availability: http://wwwdev.ebi.ac.uk/fg/sym

Acknowledgements: This project is funded by the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n°258068; EU-FP7-Systems Microscopy NoE.

BIOUML: MODULAR MODELING OF COMPLEX BIOLOGICAL SYSTEMS

Kiselev I.N.*, Kolpakov F.A.
Institute of Systems Biology, Ltd, Novosibirsk, Russia;
Design Technological Institute of Digital Techniques, SB RAS, Novosibirsk, Russia
e-mail: axec@systemsbiology.ru
** Corresponding author*

Key words: *modular approach, flattening, agent-based modeling, BioUML*

Motivation and Aim: Modeling of complex biological systems (cells, organs) is not a trivial task because of their sophisticated structure. At the same time there are a large number of different models describing particular subsystems using different formalisms and scales. Modular approach facilitates combination of such models and development of complex hierarchic models of biological systems. It has been rapidly developing in the last years [1, 2]. The aim of the present work was to develop a convenient tool for the visual creation and simulation of modular models of biological systems.

Methods and Algorithms: We have developed a plugin for BioUML open source Java framework for formal description and visual modeling of biological systems. Plugin includes graphic notation for visual creation of modular models of the biological systems. Modular models are created by connecting modules (submodels) inputs and outputs. Depending on modules formalisms different approaches are applied to the numerical simulations. In the case when all modules have the same formalisms (algebraic-differential equations with discreet events) modular model is converted into flat model appropriate for simulation. In the case of different module formalism (ODE, PDE, stochastic, etc.) modular model is simulated using agent-based approach: each module is simulated independently interacting with the others; simulation process is controlled by scheduler.

Results: Developed software provides the possibility for visual creation and numerical simulation of the modular models using special flattening algorithm and agent-modeling principles. Using the developed plugin several models have been already created:

1. The classic blood circulation model by Guyton et al. [3] (reconstructed).
2. The apoptosis model comprising 13 modules [4].
3. The complex model of the human cardiovascular system.

Conclusion: The developed software provides user with a set of tools for visual building of modular models and their simulation.

Availability: The developed software and all described models are freely available as the parts of the BioUML platform at http://www.biouml.org.

References:

1. R. Randhawa et al. (2010) Model Composition for Macromolecular Regulatory Networks, *IEEE/ACM Trans. Comput Biol Bioinform*, **7(2)**: 278-287.
2. A. Hernandez et al. (2009) A Multiformalism and Multiresolution Modelling Environment: Application to the cardiovascular system and its regulation, *Phil Trans R Soc*, **367(1908)**: 4923-4940.
3. A. Guyton et al. (1972) Circulation: Overall Regulation. *Ann Rev Physiol,* **41**: 13-41
4. E. Kutumova et al. (2012) A modular model of the apoptosis machinery, *Adv Exp Med Biol,* **736**: 235-45.

EVOLUTIONARY CONCERVATION OF NUCLEAR PORE ORGANIZATION AND COMPOSITION

Kiseleva E.V.*[1], Fiserova J.[2], Goldberg M.W. [2]
[1] *Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia;*
[2] *Durham University, UK*
e-mail: elka@bionet.nsc.ru
** Corresponding author*

Key words: *nuclear pore, nucleoporins, nuclear pore individual components, scanning electron microscopy*

Motivation and Aim: Nuclear pore complex (NPC) is the largest eukaryotic organelle which provides molecular exchange between the nucleus and cytoplasm. High resolution scanning electron microscopy let to identify many individual NPC components however only few organisms were analysed up to now. Aim of our investigations was to compare the NPC morphology from different Metazoan and Protozoa species.

Methods and Algorithms: Manual isolation of nuclei from cell or cell homogenization has been used for preparation of nuclear envelope samples. NPC structure was analysed with Field Emission in lens Scanning Electron Microscope (Hitachi, Japan).

Results: The structure of NPC individual components from *Xenopus laevis* oocytes, onion root cells, tobacco cell culture, salivary gland of *Chironomus*, yeast, *Amoeba proteus* and *Paramecium caudatum* was investigated and compared. Conservation of 8-fold symmetry as well as very high homology of individual components structure in NPCs from high animals and plants have been demonstrated for the first time. In spite that diameter of yeast NPC is smaller than those in high eukaryotes; they both are composed of a similar set of individual components. Morphology of *Amoeba* NPC at the cytoplasmic and nucleoplasmic sides of the nuclear envelope is identical to those observed in Metazoan cells despite that *Amoebae* do not have any lamina proteins. A mistake in assembly of peripheral NPC compartments in *Paramecium* was found. The basket structures instead of cytoplasmic ring with particles have been often observed at the cytoplasmic side of the nuclear envelope.

Conclusion: Comparative analysis of fine NPCs morphology from six Metazoan and two animal-like Protist species has proved the universal organization of NPC individual components in evolution. At the same time the process of peripheral NPC compartments formation in Protists is appear unstable and can develop the wrong way.

Acknowledgements: Work was supported with grants from RFBR and MCB RAS.

DISTRIBUTED ATLAS: A RULE-BASED SYSTEM FOR QUERY FEDERATION OVER SEMANTICALLY ALIGNED GENE EXPRESSION DATA SOURCES

Klebanov A.*[1, 2], Burdett T.[1], Kapushesky M.[1]
[1] *Functional Genomics Group, EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK;*
[2] *Research Centre for Technologies of Programming and Artificial Intelligence, National Research University of Information Technologies, Mechanics and Optics, St Petersburg, Russia*
e-mail: klebanov@ebi.ac.uk
** Corresponding author*

Key words: *gene expression, data integration*

Motivation and Aim: ArrayExpress Gene Expression Atlas [1, 2] is an added-value database of gene expression statistical analysis based on expert-curated public data. With Atlas software available for standalone installations, several instances in different institutions exist today serving specialized needs: for private data sets within enterprise firewalls, hosting new data generated in research consortia or custom-curated data sets for project-specific uses. This work addresses the need of researchers using the Atlas technology to perform complex queries across several Atlases simultaneously.

Methods and Algorithms: Distributed Atlas uses a client-server architecture with a single client aware of multiple independent servers. Compared to the original Atlas design it represents a significant architectural evolution resulting in a more flexible infrastructure built with reusable components.

A distributed query is a query (query types are described in [1]) performed on a set of registered data sources. Given $n \geq 2$ sources a conflict arises: each query returns n results, when normally just one result is required. A gene search, for instance, queries several sources to find out whether the gene has been differentially expressed in experiments stored in them. A number of same genes but with different expression values will be retrieved forcing researchers to analyse them manually in order to gain a generalized result. Such analysis becomes infeasible with a number of genes and data sources increasing leading to the need of an automatic solution. To tackle this problem several (distributed) query federation rules have been introduced.

A query federation rule is a mapping of an n-dimensional query result vector onto a single entity of the same domain. Formally, let D be a single query result domain, $r = (r_1, \ldots, r_n)$ be an n-dimensional federated query result vector, where $r_i \in D$. Thus, a query federation rule (denoted by qfr) is defined as qfr: $D_n \rightarrow D$.

We designed the following distributed query federation rules:

'First not null' rule. Starting from k = 1 and going to n, if the response from k-th server is not null or empty, then this response is used as the query result. The intuition here is to return the first meaningful result.

'Aggregate' rule. All n responses are combined into a single multi-source result object, discarding exact duplicates. In some cases this rule is insufficient and additional recalculations are performed.

Conclusion: We demonstrated that Distributed Atlas provides a powerful interface for federated querying of multiple semantically aligned data sources. While the public Gene Expression Atlas at EBI is itself a powerful resource for the general researcher, we developed a means for multiple instances of standalone Atlas software with additional expression data to be easily integrated in a meaningful manner, without compromising on Atlas query power or Atlas user interface capability.

Availability: The software is available at http://www.ebi.ac.uk/fg/gxa-distributed.

Acknowledgements: We thank Alvis Brazma for valuable comments.

Funding: This work was supported by the European Commission [226073, 242220].

References:

1. M. Kapushesky, et al. (2010). Gene Expression Atlas at the European Bioinformatics Institute. *Nucl. Acids Res.*, **38**: D690–D698.
2. M. Kapushesky, et al. (2012). Gene Expression Atlas update – a value-added database of microarray and sequencing-based functional genomics experiments. *Nucl. Acids Res.*, 40: D1077–D1081.

THE LARGE-SCALE ANALYSIS OF *ARABIDOPSIS THALIANA* MUTANT *LEL* USING RNA-SEQ

Klepikova A.V. *[1, 3], Logacheva M.D. [2, 3], Penin A.A. [1, 3]
[1] *Department of Genetics, Biological Faculty, M.V. Lomonosov Moscow State University, Moscow, Russia;*
[2] *Department of Evolutionary Biochemistry, A.N. Belozersky Institute of Physico-Chemical Biology, M.V. Lomonosov Moscow State University, Moscow, Russia;*
[3] *Evolutionary Genomics Laboratory, Faculty of Bioengineering and Bioinformatics, M.V. Lomonosov Moscow State University, Moscow, Russia*
e-mail: annklepikova@gmail.com
** Corresponding author*

Key words: *RNA-seq, Arabidopsis thaliana, development, meristem*

Motivation and Aim: Arabidopsis thaliana mutant *lel* demonstrate the loss of petals and stamens. Primary hypothesis suggest involvement of gene *LEL* in maintenance of floral meristem. We use RNA-seq to identify the place of *LEL* in genetic network controlling meristem development and maintenance and to investigate changes in gene expression caused by mutation in gene *LEL*.

Methods and Algorithms: Plant material from wild type and mutant plants was collected in two developmental stages: apical meristem when flower primordial are formed and young inflorescences on the stage of the anthesis of the first flower in two biological replicates. Two independent RNA-seq experiments were performed using sequencing on Illumina HiSeq2000. Results of RNA-seq were analyzed using CLC Genomics Workbench 5.0.1 software.

Results: We found about 20000 expressed genes in both wild type and mutant and about 1500 of them were differentially expressed. The majority of them controlled response to different stresses and metabolic pathways. Also we observe difference between wild type and mutant in expression of genes controlling meristem maintenance: expression of the main stem cell activity regulator *WUSCHEL* decreased five times, when its negative regulator *CLAVATA1* increased by half.

Conclusion: The profile of gene expression changes allows us to define *LEL* place in regulatory network as novel regulator of meristem maintenance.

The study is supported by the Russian Foundation for Basic Research (project № 12-04-01599).

HAPLOID EVOLUTIONARY CONSTRUCTOR: A GRAPHICAL USER INTERFACE FOR SIMULATING BACTERIAL COMMUNITIES EVOLUTION

Klimenko A.I., 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: *computer simulations, modeling, evolution, evolutionary constructor*

Motivation and Aim: Computer simulation and modeling is one of the base tools used for the theoretical investigation of evolution. Proper graphical representation of simulation results sufficiently facilitates the model analysis and allows a researcher to do the best interpretations of simulations. The control usability of a model also accelerates scientific research. This study is devoted to the development of several graphical user interface (GUI) program components for the software package "Haploid evolutionary constructor" (HEC) [1]. This software is used for the computer simulation of population and evolutionary processes, occurring in bacterial communities. There is a wide range of biological processes modeled in HEC: mutation, horizontal gene transfer and gene losses, phage infestation and speciation etc. It is important to note, that the HEC model simultaneously describes several layers of biological organization: genetic, metabolic, population, ecological ones [2]. Summarizing all the above mentioned, it is clear that GUI should be able to visualize data of various types, as well as GUI should give a user possibility to control every parameter of such a complex model.

Methods and Algorithms: We used the Model-View-Controller (MVC) architecture and separated three components in order to minimize the mutual influence of the data model, graphical interface and user control. As the computational core of HEC was written in C++ and the GUI was written in Java, we used the Java Native Interface (JNI) technology to connect them. The GUI framework was written using Swing, and the JFreeChart library (http://www.jfree.org/jfreechart/) was used for plots visualization. The visualization of graphs was performed by the JGraphX library (http://www.jgraph.com/jgraph.html).

Results and Conclusion: We have developed the GUI components for the HEC software package, which provide convenient visualization of HEC data, model construction, setting up and control. The GUI also provides the model persistence by the generation of model scripts using special HEC language. It is a convenient tool for simulation of bacterial communities evolution.

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

Acknowledgements: The work was supported by the RFBR grants 10-04-01310-a, 12-07-00671-a, Interdisciplinary integration projects of SB RAS № 47, 130.

References:

1. S.A. Lashin et al. (2010) Comparative modeling of coevolution in communities of unicellular organisms: adaptability and biodiversity, *JBCB*, **8:** 627-643.
2. S.A. Lashin et al. (2011) Trends in the Prokaryotic Community and Prokaryotic Community–Phage Systems, *Russian journal of genetics,* **47:** 1487-1495.

INFERENCE OF SIGNALING NETWORKS USING A LINEAR MODEL

Knapp B.*[1,2], Mazur J.[1,2], Kaderali L.[1,2]
[1] *Heidelberg University, ViroQuant Research Group Modeling, BioQuant BQ26, Im Neuenheimer Feld 267, 69120 Heidelberg, Germany;*
[2] *Dresden University of Technology, Medical Faculty Carl Gustav Carus, Institute for Medical Informatics and Biometry, Fetscherstrasse 74, 01307 Dresden, Germany*
e-mail: bettina.knapp@tu-dresden.de
** Corresponding author*

Key words: *RNA interference, network inference, linear model*

Motivation and Aim: RNA interference (RNAi) is a powerful tool to identify gene function and gene involvement in biological processes. Recent technical developments allow facilitated measurements of the data and thus, qualitative and quantitative improvements of high-content and high-throughput experiments. However, learning the underlying network of a biological process with RNAi data remains a challenging task. One of the problems is the dimensionality which increases exponentially with the network size. Furthermore, the given data is in most cases noisy, incomplete or both.

Methods and Algorithms: We present a model which infers signaling networks using a linear optimization program which can be solved efficiently even for larger network sizes. The model can easily deal with double or multiple knockdowns and integrate prior knowledge information.

Results: Based on data simulated for networks of different sizes we show that our method is better than random guessing and that it outperforms a recently published approach, especially when applied to large-scale problems. We use different levels of noise and missing data to show that the model can deal with incomplete and noisy data. Furthermore, we achieve a significant reduction in computation time in comparison to the other approach. Using real biological data studying the ERBB signaling pathway we could confirm several already known gene interactions as well as identify potential new ones. In total, the accuracy, computed based on the STRING database, of the ERBB network learned with our model is better than random guessing.

Conclusion: Formulating the problem of network inference as a linear model allows a fast and efficient computation of the underlying topology. This may help to get a better understanding of the underlying biology and thus, to identify new drug targets in the future.

Availability: The R source code of the method is available on request from the authors.

Acknowledgements: We acknowledge funding from the German Ministry of Education and Research (BMBF) via the ForSys/ViroQuant programs (grant0313923), and the European Union seventh framework program via the PathoSys project (grant number 260429).

EXPERIENCE OF THE PERSONALIZED ANTIPLATELET THERAPY: THE EFFECTS OF CYP2C19 GENE

Knauer N.Yu.*, Voronina E.N., Lifshits G.I.
Novosibirsk State University, Novosibirsk, Russia;
Institute of Chemical Biology and Fundamental Medicine SB RAS, Novosibirsk, Russia
e-mail: knauern@mail.ru
** Corresponding author*

Key words: *clopidogrel, personalized therapy, CYP2C19*

Motivation and Aim: The clopidogrel-based prevention of thromboses and rethromboses is the important aspect of cardiological practice. However, clinical and laboratorial response on clopidogrel of different patients is known to differ from the predicted one thus stipulating the use of personalized approach for search of the optimal dose. One of the principal instruments of this approach is the revelation of the allelic variants of genes influencing the drug metabolism. Particularly, the protein CYP2C19 of the cytochrome P450 family was shown to play the important role in the clopidogrel metabolism, with its activity depending of the allelic variants of the gene.

The aim of this work was to examine frequencies of occurrence of CYP2C19 allelic variants *1, *2, *3 among the patients receiving clopidogrel by medical indications (n=158) and to determine their contribution to the clopidogrel laboratorial efficacy

Methods and Algorithms: The assessment of the laboratorial response on clopidogrel was conducted by light transmission aggregometry using ADP (20 μM). The identification of the CYP2C19 allelic variants was performed by real-time PCR-HRM assay using specific primers and also by RFLP assay. The results were compared with the aggregometry data to elucidate the effect of polymorfisms on the platelet aggregation after the clopidogrel taking.

Results: Depending on the platelet aggregation change after clopidogrel taking, the following groups were assigned: responders (58.9%), semi-responders (20.9%) and non-responders (7.6%). In addition, the group of patients (12.7%) was assigned, with the platelet aggregation increasing up to 3.5 times after the taking of clopidogrel. According to the data of CYP2C19 allelic variants identification, the whole group of patients contained allelic variants CYP2C19*1 (74.7 %), CYP2C19*2 (24.7%), CYP2C19*3 (0.6%). The response on clopidogrel was shown to correlate with the allelic variant CYP2C19*2 and not to correlate with the variant CYP2C19*3.

Conclusion: The results obtained are of the practical interest and can be used for the optimization of the clopidogrel-based antiplatelet therapy.

DEVELOPMENT OF THE SOFTWARE COMPLEX "GENETICS" FOR SUPPORT INVESTIGATIONS IN MEDIC GENETICS

Kolpakov F.A.[1, 2*], Tyazhev I.[1], Tolstykh N.[1], Kudryavtseva E.A.[3], Sharipov R.N.[1, 4], Boyarskikh U.[3], Kondrakhin Yu.[1, 2], Filipenko M.L.[3], Lifshits G. [5]

[1] *Institute of Systems Biology, Ltd, Novosibirsk, Russia;*

[2] *Design Technological Institute of Digital Techniques, SB RAS, Novosibirsk, Russia;*

[3] *Institute of Chemical Biology and Fundamental Medicine, SB RAS, Novosibirsk, Russia;*

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

[5] *Center of New Medic Technologies, SB RAS, Novosibirsk, Russia*

e-mail: fedor@biouml.org

** Corresponding author*

Key words: *software complex "Genetics", medic genetics research, support of complex investigations, statistical processing of data*

Motivation and Aim: Modern medic investigations demands support by advanced software complexes providing data storage, tools for mathematical processing of data and secure sharing of data for collaborative work worldwide. Often several medical centers are involved to collect representative samples, perform statistical compilations and interpret the obtained results. In this case, application of a comprehensive software complex supporting joint remote work and data processing in the frames of the single platform is indispensable and will bring more fruitful results than the separate work. Hence, development of such software complexes is one of the key tasks of the modern applied sciences.

Methods and Algorithms: The complex "Genetics" is based on a classic 3-tier architecture: 1)MySQL database, 2)application server (business logics - BeanExplorer EE, http://www.beanexplorer.com; statistical calculations – BioUML (http://www.biouml.org) and R/Bioconductor (http://www.bioconductor.org/), and 3)GUI (Internet browser).

Results: "Genetics" was developed with application of the advanced web technologies as a result of collaboration of the SB RAS institutes. The complex allows comprehensive work with patient data starting from medic questionnaires (>200 parameters) to data set building and statistical data processing (e.g., genetic polymorphisms). Special attention was paid to secure access of the authorized personnel only to detailed data on patients (1st – 3rd category [1]) taking into account the Russian federal laws on personal information protection. For calculations depersonalized data (4th category) are used that allows to involve specialists from other centers or countries. The complex was deployed on a special server reachable via http protocol and provides full spectrum of analysis methods (both self-developed and available via R/Bioconductor). The complex was tested successfully and is used now by the Center of New Medic Technologies SB RAS.

Availability: http://genetics.biouml.org/genetics/

Acknowledgements: This work was supported by the Presidium of the Russian Academy of Sciences (The "Basic Science for Medicine" Program, grant Nr.33).

References:

1. The Russian Federal Law №152 as of 27.07.2006. «On Personal Data» http://base.consultant.ru/cons/cgi/online.cgi?req=doc;base=LAW;n=117587

ALTERNATIVE HYDROGEN BONDING IN MOLECULAR DESIGN OF THERMOSTABLE ANTIOXIDANT PROTEIN

Kondratyev M.S.*, Kabanov A.V., Novoselov V.I., Samchenko A.A., Komarov V.M., Khechinashvili N.N.
Institute of Cell Biophysics RAS, Pushchino, Moscow region, Russian Federation
** Corresponding author: e-mail: ma-ko@bk.ru*

Key words: *protein structure, thermostability, molecular dynamics, peroxiredoxin VI, amino acid substitutions, homology*

Motivation and Aim: On of the goal of modern bioengineering is concerned with the development of molecules with predefined properties. Mainly it concerns the increase in stability of macromolecules, including thermostability of enzymes. That allows to speed-up of biocatalysis and protection the macromolecule from unfolding. Our work use newest theory of thermostabilization of small globular proteins [1, 2], developed in the our Laboratory of Structure and Dynamics of Biomolecular Systems ICB RAS. It is based on the alternative hydrogen bonding between side chains of amino acids on protein surface. Object of our investigations was human Peroxiredoxin VI (PRX6) – new promising antioxidant for burn treatment, discovered and described in our Institute [3, 4].

Methods and algorithms: homology analysis, molecular dynamics (GROMACS)

Results: analysis of homology, known spatial structure PRX6 and calorimetric data [4] give us information about localization and importance amino acid substitutions in human PRX6. Evolutionary changes of this protein lead us to search the most probable sites of mutations only in variable areas of amino acid sequence, because changes in stable regions can affect the functional properties of this antioxidant enzyme. It is important to notice that human and rat PRX6 have the highest homology (91,5 %, i.e. 19 residues) and rat protein possesses the greatest thermostability of studied PRX's. The structures of native human Peroxiredoxin 6 and it homologs has been studied by long full-atomic molecular dynamics simulation with solvent at various temperatures on GPU NVIDIA. We counted the amount of hydrogen bonds (salt bridges) on the surface of protein globules in each frame of an MD-trajectory and propose only 4 amino acid substitutions (V10E, N107D, I165D, D183K), which we believe will lead to increased thermostability of human PRX6 without distortion of the 3D-structure and antioxidant activity.

References:

1. Khechinashvili N.N., Volchkov S.A., Kabanov A.V., Barone G., Thermal stability of proteins does not correlate with the energy of intramolecular interactions. // Biochim. Biophys. Acta (BBA) Proteins & Proteomics 2008, 1784 (11), P.1830.
2. Khechinashvili N.N., Fedorov M.V., Kabanov A.V., Monti S., Ghio C., Soda K., Side Chain Dynamics and Alternative Hydrogen Bonding in the Mechanism of Protein Thermostabilization. // J. Biomol. Struct. & Dyn., 2006, 24(3), P.255-262.
3. Peshenko I.V., Novoselov V.I., Evdokimov V.A., Nikolaev Yu.V., Shuvaeva T.M., Lipkin V.M., Fesenko E.E. // Novel 28-kDa secretory protein from rat olfactory epithelium. FEBS Lett. 1996, 381, P.14-19.
4. Sharapov M.G., Novoselov V.I., Ravin V.K. Cloning, expression and comparative analysis of peroxiredoxine 6 from different species // Mol Biol (Mosk). 2009, May-Jun; 43(3). P.505-11.

BIOINFORMATICS APPROACH TO THE STUDY OF DYSTROPIC DISEASES

Koneva L.A.*[1], Bragina E.Yu.[1], Tiys E.S.[2], Freidin M.B.[1], Ivanisenko V.A.[2], Puzyrev V.P.[1]

[1] *Research Institute of Medical Genetics, SB RAMS, Tomsk, Russia*
[2] *Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia*
e-mail: lada.koneva@medgenetics.ru
** Corresponding author*

Motivation and Aim: The concept of dystropy describes rare combination of diseases in same individual and Amutual repulsion@ of clinical phenotypes (1, 2). We sought to investigate this phenomenon in the study of genetic relationships between bronchial asthma (BA) and tuberculosis (TB) as a well-documented example of a dystropy. We assume that rare co-occurrence of these diseases is caused by the genes involved in the pathophysiological mechanisms that arrest simultaneous manifestation of BA and TB. Identification of common and specific genes is important for determine the molecular genetic mechanisms leading to uncommon co-occurrence of these diseases in same individual. We carried out a search and analysis of common genes underlying BA and TB to develop a strategy for dystropic disease study; then we plan to focus on disease specific genes.

Methods and Algorithms: We started our analysis with the identification of proteins that were associated both with BA and TB. We reconstructed two associative networks, including all the proteins associated with the diseases of interest. We used ANDCell (3), a computer-based information system for automated retrieval and integration of associative knowledge from factual and text sources of knowledge. The BA network includes 561 proteins and the TB network 129 proteins. We detected 66 Ashared@ proteins associated with the development of both BA and TB by comparing the two lists of proteins of the associative networks. For these proteins was carried out a review to confirm that the proteins in the network are actually associated with two diseases of interests. The list of 66 "shared" proteins was reduced to 22. Then, we checked whether genes encoding for 22 "shared" proteins are associated with BA and TB using the information resource HuGE Navigator which allow performing a search in the Human Genome Epidemiology Network (HuGENet).

Results: After the review of the "shared" genes, we retained 10 already studied genes associated with both BA and TB and three previously unstudied genes (*IFNA1, HP,* and *CD79A*), which can be considered as novel candidate genes of these diseases. The analysis of the 10 previously established genes and their proteins revealed that these proteins are not specifically involved in the pathogenesis of BA and TB; rather, they only feature general immune processes such as inflammation.

References:

1. Pfaundler M., von Seht L. (1921) Weiteres uber Syntropie kindlicher Krankheitzustande. *Z Kinderheilkd*, **30**: 298-313.
2. V.P. Puzyrev, M.B. Freidin (2009) Genetic View on the Phenomenon of Combined Diseases in Man, *Acta Naturae*, **3**: 6-11.
3. P.S. Demenkov et al. (2008) Associative network discovery (AND) B the computer system for automated reconstruction networks of associative knowledge about molecular-genetic interactions, *Computational technologies,* **13:** 15-19.

COMBINATION OF PROTEIN-PROTEIN INTERACTION NETWORK ANALYSIS AND DISCRETE MODELING FOR IDENTIFICATION OF PROMISING PHARMACOLOGICAL TARGETS FOR ALZHEIMER'S DISEASE

Konova V.I.*, Koborova O.N., Filimonov D.A., Poroikov V.V.
Orekhovich Institute of Biomedical Chemistry of RAMS, Moscow, Russia
e-mail: varvara.konova@ibmc.msk.ru
** Corresponding author*

Key words: *Alzheimer's disease; network analysis; drug target identification; protein–protein interactions*

Motivation and Aim: The disbalance of regulatory components in pro- and antiapoptotic systems of the cerebral cortex neurons leads to the perturbation of cell cycle and neuronal death during aging and development of neurodegenerative diseases. The aim of the research is to find key proteins involved in development and progression of Alzheimer's disease (AD) and identify the most prospective pharmacological targets for neurodegeneration treatment.

Methods and Algorithms: We developed an algorithm for drug target identification based on discrete modeling of regulatory networks and applied it for analysis of AD network. The algorithm models cell cycle regulation as dichotomic networks using regulatory network on the cellular signaling processes (data from the TRANSPATH™ database (http://www.biobase.de) and literature) and the microarray data of proteins and/or genes as a primary network state [1]. Protein-protein interaction (PPI) network analysis was performed by Cytoscape plug-in APID2NET [2].

Results: The method was applied to AD cell cycle regulatory network consisted of more than 1000 nodes and edges. To reveal pathological changes in the regulatory network we used the gene expression data on patients with AD compared to the normal aged patients (ArrayExpress database http://www.ebi.ac.uk/arrayexpress/) and found both known and new pharmacological targets for AD. PPI analysis identified 129 upregulated proteins (hubs) playing key role in AD. As a result, the intersection of targets' lists obtained by dichotomic modeling and PPI analysis revealed the most promising targets for treatment of AD.

Conclusion: Combination of standard PPI analysis and discrete dichotomic modeling method demonstrated the applicability for drug targets identification for AD. As a result new pharmacological targets were found that have to be further validated experimentally.

Availability: by request

References:

1. O.N.Koborova et al. (2009) In silico method for identification of promising anticancer drug targets, *SAR QSAR Environ Res.*, **20**: 755-66.
2. J.Hernandez-Toro et al. (2007) APID2NET: unified interactome graphic analyzer. *Bioinformatics,* **23**: 2495-2497.

LINGUISTIC ANALYSIS OF SHORT SEQUENCES IN THE INTRONS AND EXONS FOR TLR1

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

Key words: *TLR genes, linguistic approach, promoters, UTRs*

Motivation and Aim: Eukaryotic genes contain all informations essential for efficient transcription and subsequent translation for forming functional proteins. Part of this information is used during transcription process and remaining part is passed on to the next level of regulation, i.e. translation. Since the cellular processes are highly energy efficient, it can be hypothesized that the part of information used in the transcription step, which is not required for succeeding steps (i.e. translation) will be 'discarded', i.e. will not be carried further. Therefore, a direct comparison of the gene with the gene product (i.e. mRNA and protein) can tell us the part of the gene that has contributed for the regulatory process. Thus, the regions denoted as vestigial can in fact turn out to be inevitable for the preceding step. Comparison of these gene segments can provide us valuable regulatory informations.

Methods and Algorithms: TLR1 gene was selected from the Ensembl database for human, mouse and zebrafish. These sequences were compared with the corresponding mRNA sequences. The remaining part of the gene were searched for regulatory regions of the alternative splicing sites. The regions that were not carried over to the next step were assumed to be regulatory regions. These regions were searched for similarities in the promoter, 3'-UTR and intron regions of the respective gene. The sequences were also compared with the mature human miRNA (obtained from http://mirbase.org/), since these are generally attributed for their regulatory action.

Results: The 6-nt sequences (339) which were found to be ubiquitously present in set of introns (from all the three species) and miRNA, were further analyzed and searched for conserved pattern. Similar tests was performed for the promoter regions and 3'-UTR regions. On comparing the sequences among the three species, a broad idea about the most commonly found sequences indicate the genome's idea of simple words. Words that are widely occuring (more frequent) carry less information compared to highly specific (less frequent) words. While comparing introns from three species and miRNA from human, it was found that there are 55 sequences which are common, but are present in the introns of all the species just once.

Detailed results of the analysis will be presented.

Conclusion: Sequences present just once in the whole gene and present in regulatory segments as miRNA can be actively involved in the regulation of the gene processes.

Acknowledgements: I thank C. K. Mitra for valuable discussions.

MODELLING KREBS CYCLE AS AN ELECTRICAL CIRCUIT

Korla K., Mitra C.K.*
University of Hyderabad, Hyderabad 500046, India
e-mail: c_mitra@yahoo.com

Key words: *Krebs cycle, electrical analogue, flux, feedback mechanism*

Motivation and Aim: Biochemical metabolic pathways contain information regarding all the important biochemical reactions within a cell. Although detailed information about the various metabolic enzymes are available, the actual computation of the various fluxes is difficult. Electrical circuits are easier to analyze and study, as several useful tools are already available for this purpose. Here, we propose an electrical circuit analogue for the well known Krebs cycle, which can be used for the determination and studies of various fluxes relatively easily.

The complete cycle is under tight regulation and is maintained by the rate of conversion of pyruvate to acetyl-CoA and by the flux through citrate synthase, isocitrate dehydrogenase, and α-ketoglutarate dehydrogenase (they are the regulatory components in the cycle). These fluxes are mainly controlled by the concentration of substrates and products, the end product as NADH, ATP, citrate show inhibitory effect and the substrates as NAD+ and ADP are stimulatory. While modelling an enzyme with an electrical analogue we note that an enzyme with a regulatory site can be modelled as an amplifier with negative feedback. The negative feedback causes the output to stay stable in spite of changes in the input signal. The negative feedback also ensures that the role of the amplifier (the specific activity of the enzyme) is minimal and the feedback elements contribute to the overall performance of the circuit.

Methods and Algorithms: A standard copy of the Krebs cycle was downloaded from wikipedia. We have modified this cycle so that enzymes appear at the nodes and the substrates appear along the edges. Within a cell, connection between different nodes is established by products and reactants. Connection between different pathways can also occur via nodes, i.e., enzymes, that may be shared between two different circuits. In an electrical circuit this connection is made by physical metallic wire that are responsible for electron conduction. This feature is absent in a cell as enzymes are highly specific and will respond only to specific substrates. A simple enzyme (without any regulatory site) simply acts like a buffer in an electrical circuit. Its only purpose is to accept a given substrate and produce a corresponding product. This product may enter another cycle or may continue within the same series of reactions. A regulatory enzyme acts more like an amplifier, whose gain can be controlled by another molecule. Such a system is best represented with an amplifier with a feedback network. The feedback produces an output which is stable in spite of variations in input signal.

Conclusion: We have modelled the Krebs cycle as a series of five amplifiers that are connected in a cyclic manner. In a biochemical reaction, unit input flux always gives rise to unit output flux and the operational amplifiers in the above scheme have gains very close to unity. This is a very good overall simplification of the Krebs cycle, that can be used in actual kinetic modelling if sufficient detailed parameters of the enzymes are available.

Simulation results will be presented based on the electrical circuit.

GENETICS AND DISEASE PROGRESSION OF FAMILIAL MULTIPLE SCLEROSIS IN NOVOSIBIRSK REGION OF RUSSIA

Korobko D.S.*[1], Malkova N.A.[2], Kudryavtseva E.A.[3, 4], Filipenko M.L. [3]
[1] State Novosibirsk Regional Clinical Hospital, Regional MS Centre; [2] Novosibirsk State Medical University, [3] Institute of Chemical Biology and Fundamental Medicine, Siberian Division, Russian Academy of Sciences; [4] Novosibirsk State University, Novosibirsk, Russia
e-mail: denn007@ngs.ru
** Corresponding author*

Key words: *multiple sclerosis, familial cases, single nucleotide polymorphisms, TNFα, recurrence risk*

Motivation: Multiple sclerosis (MS) is a chronic disorder of the central nervous system, characterized by inflammation, demyelination, development of plaque lesions and episodes of neurologic dysfunction that frequently leads to progressive degeneration [1]. MS is a multifactorial disease in which both genetic and environmental factors intervene. In some studies it is observed that familial multiple sclerosis has more benign course in compared to sporadic cases. For neurologists it is important to predict course of such heterogeneous disease as MS.

Aim: To investigate clinical and genetic features of familial MS.

Materials and methods: It is ongoing retrospective-prospective study. Up to date 248 patients with MS according to the diagnostic criteria of McDonald et al. (2005) were recruited to study in Novosibirsk Regional MS Center. Single nucleotide polymorphisms (SNPs) in *CD40* (rs6074022, rs1883832, rs1535045, rs11086998), *TNF-a* (rs1800629) and *TNFRSF1A* (rs4149584) and rs3135388 (genetic marker of allele HLA-DRB1*15) were genotyped by real time PCR, using competing TaqMan probes. The control group (n=567) comprised of people without inflammatory CNS disease living in Novosibirsk. The study was approved by local ethics committees and all participants signed informed consent forms. To assess disease progression MS severity score (MSSS) [2] and rate of progression [3] was calculated. MSSS was determined in 238 patients with disease duration > 1 year.

Results: 17 familial cases with relapsing-remitting MS who belonged to 15 different families were identified. Recurrence risk in families of this group was 6,0 %, total recurrence risk was calculated (6,9 %). Age at MS onset in familial cases was lower than in sporadic (23,03 ± 9,22 vs 27,28 ± 8,59, p = 0,09). The rate of disease progression and MSSS within family group was significantly lower (0,33 vs 0,69, p < 0,001; MSSS: 3,22 vs 4,4, pANOVA = 0,05). The association between SNP rs3135388 and development of MS was found (OR = 3.04, 95 %CI 2.33-3.96, p = $1.7*10^{-17}$), but this polymorphism has not influence on disease progression. The association between *GA*, *AA* TNF*a* genotypes and higher average annual relapse rate was revealed (OR 6,07, 95 %CI 1,19-30,9, p = 0.032). These genotypes were more frequent (23 % vs 10 %) in subjects without family history of MS.

Conclusion: This study of Siberian cohort confirmed more benign course of familial MS. The further research is needed to understanding the genetic basis of susceptibility in MS.

References:

1. Compston A, Coles A. Multiple sclerosis. Lancet. 2008 Oct 25; 372(9648):1502-17.
2. Roxburgh RHSR, Seaman SR, Masterman T. et al. Multiple Sclerosis Severity Score: Using disability and disease duration to rate disease severity Neurology April 12, 2005 64:1144-1151
3. Verjans E., Theys P., Delmotte P. et al. Clinical parameters and intrathecal IgG synthesis as prognostic features in multiple sclerosis. Part I. J Neurol. 1983; 229(3): 155-65.

MS/MS ANALYSIS OF METABOLIC DISORDERS

Koval V.V.*, Alekseeva I.V., Chernonosov A.A., Fedorova O.S.
Institute of Chemical Biology and Fundamental Medicine, SB RAS, Novosibirsk, Russia
e-mail: koval@niboch.nsc.ru
** Corresponding author*

Key words: *metabolic disorders, tandem mass-spectrometry*

Motivation and Aim. Metabolic disorders are typically accompanied by the accumulation of corresponding metabolites in blood. Using modern tandem mass spectrometry (MS/MS) more than 20 inherited metabolic disorders can be detected simultaneously from a single blood spot by quantifying concentrations of up to 50 metabolites. Many disorders of fatty acid oxidation, organic acid metabolism, and amino acid metabolism can be detectable by use of MS/MS. The use of MS/MS to detect metabolic disorders in newborns is one of the most important advancements in neonatal screening. In spite of errors of metabolism are a rare group of genetic disorders (0.001–0.01 %) they can produce serious clinical consequences. Early medical intervention in many cases provides the only possibility to avoid physical and mental retardation, or even death.

Methods. The sample preparation utilizes ultrasonic extraction of the blood spot from paper (Whatman 903, Whatman Inc., NJ, USA) in methanol. The extraction is followed by butyl-esterification of the free carboxyl groups of amino acids (requiring hydrochloric acid plus butanol). Esterification converts zwitterionic amino acids to amino esters similar to simple amines, dramatically enhancing the ionization efficiency in triple-quadrupole mass spectrometers Agilent 6310 Triple Quard LC/MS (Agilent Technologies, USA). Most amino acid butyl esters show a characteristic loss of neutral butyl-formate (102 Da) in collision-induced dissociation (CID). Monitoring this process by the constant neutral loss scan technique allows simultaneous detection of these compounds.

Results. The blood-spot samples from 255 patients of Novosibirsk clinics, including inborn, were analyzed. Phenylketonuria (PKU) was confirmed in 28 cases. Besides, 18 cases of urea cycle disorders, 104 cases of short- and medium-chain fatty acid catabolism defects, 12 cases of homocysteine/methionine synthesis defects were detected. Often the diseases have complex character and represent the superposition of few disorders.

Acknowledgements. Supports by grants from SB RAS (no. 5), RAS "Fundamental sciences to medicine" (no. 5.6 and Ministry of Education and Sciences (no. 16.512.11.2073).

MASS-SPECTROMETRY-BASED IDENTIFICATION OF ENDOGENOUS PEPTIDES IN BLOOD SERUM

Kovalchuk S.I.*, Ziganshin R.H., Arapidi G.P., Azarkin I.V., Govorun V.M., Ivanov V.T.
Institute of Bioorganic Chemistry RAS, Moscow, Russia
e-mail: xerx222@gmail.com
** Corresponding author*

Key words: *endogenous peptides, mass-spectrometry, serum*

Motivation and Aim: In 2011 the Human Proteome project was started, however most of the well-known and reliable protocols (both experimental and computational) are designed to study large protein molecules. The classic way to study proteome is, for example, to make 1D or 2D electrophoresis, to treat different parts of a gel containing moderate number of proteins with trypsin and to analyse the tryptic peptide fragments by LC-MS/MS. Another way is to use different immunochemical approaches. However both these approaches miss such a huge group of small proteins as peptides being the products of natural degradation of the host proteins while peptidome studies could give invaluable data concerning internal human body processes. Most of the present papers on human peptidomics work with body liquids such as saliva, cerebrospinal and synovial liquids, urea, and most of all with blood serum and plasma. And this is the place where the troubles begin: nonpeptide contaminants (proteins, hydrocarbons, low molecular weight impurities etc.) impeding normal chromatographic peptide separation, incredible native peptide diversity with very high concentration dynamic range… The only known way to study peptidomes consists of prefractionation of native peptide mixtures with subsequent analysis by high-mass-accurate LC-MS/MS with modern fast and precision mass-spectrometers such as Q-TOF, Orbitrap and TripleQuad machines.

Methods and Algorithms: Our group works on the human serum peptidome. As a first step we separate serum samples on magnetic beads with weak cationic exchange surface followed by several alternative separation methods: SAX-HPLC, RP-HPLC and IEF. The resultant samples are analysed by RP-LC-MS/MS using Q-TOF mass-spectrometer and the mass-lists are searched against SwissProt database.

Results: We also add one special step while making samples – we heat the eluates after magnetic beads separation at 95^{o}C for 15 min. This leads to peptide desorption from the major serum proteins particularly from albumin. This step utterly increases the number of analyzable peptides – based on the LC-MS analysis this number rises up to many thousand individual compounds. This quantity clearly demonstrates the necessity of the fast and sensitive MS equipment to analyse more peptides beyond the most abundant ones. For now we have gathered a database containing about 3000 thousand unique peptides which significantly surpasses all the published results.

Conclusion: However it is important to understand that exhaustive analysis of peptidome cannot be an end in itself. The serum peptide database we are creating contains not only peptides from healthy donors, but also from the people with different socially-significant diseases and can be used for potential peptide biomarker search. Besides, the knowledge about changes in degradome activity could be of great importance for studying molecular mechanisms of different human diseases developing.

DIRECT COMPUTER SIMULATION OF PROTEIN-PROTEIN INTERACTION

Kovalenko I.B.
Biological faculty of Moscow Lomonosov State University, Moscow, Russia
e-mail: ikovalenko78@gmail.com

Key words: *Brownian dynamics, diffusion, electrostatic interaction, protein*

Motivation and Aim: The goal of this work is to study the kinetics of protein-protein interactions between the electron transport proteins involved in photosynthesis by means of computer simulation.

Methods and Algorithms: We present a new method for computer simulation of formation of protein-protein complexes in a cell environment (1). The method makes it possible to simulate association reactions of several hundreds of protein pairs in sub-cellular compartments, and to obtain the real-time dynamics of protein-protein interactions. The method allows us to explore the effect of electrostatic forces on the protein-protein complex formation and evaluate the kinetic rate constants.

Results: Calculations correctly reproduce binding interactions between the electron transport proteins involved in photosynthesis for different values of ionic strength in the solution and in the chloroplast thylakoid lumen, while taking into account electrostatic interactions between proteins and the thylakoid membrane (2, 3). The model demonstrates non-monotonic dependences of complex formation rates on the ionic strength as the result of long-range electrostatic interactions (4). The developed method can also be used as a predictive tool to resolve the binding sites and to describe complex structures for a range of proteins (5).

Conclusion: The simulation method presented in this work serve to reveal the molecular interactions (diffusion, electrostatic interactions) underlying the arrangement of photosynthetic electron transport regulation.

Availability: Available on request from the authors.

Acknowledgements. The work is supported by grants from the Russian Foundation of Basic Research (11-04-01019 and 11-04-01268).

References:

1. I.B. Kovalenko et al. (2006) Direct simulation of plastocyanin and cytochrome f interactions in solution, *Phys. Biol.* **3:** 121-129.
2. I.B. Kovalenko et al. (2010) Direct computer simulation of ferredoxin and FNR complex formation in solution, *Phys. Biol.* **7:** 26001.
3. O.S. Knyazeva et al. (2010) Multiparticle computer simulation of plastocyanin diffusion and interaction with cytochrome f in the electrostatic field of the thylakoid membrane, *Biophysics* **55:** 221-227.
4. I.B. Kovalenko et al. (2011) Computer simulation of interaction of photosystem 1 with plastocyanin and ferredoxin, *BioSystems* **103:** 180-187.
5. I.B. Kovalenko et al. (2009) A novel approach to computer simulation of protein–protein complex formation, *Dokl. Biochem. Biophys.* **427:** 215–217.

IMPROVED DIFFERENTIAL EVOLUTION ENTIRELY PARALLEL METHOD

Kozlov K.N.*[1], Samsonov A.M.[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: kozlov@spbcas.ru
** Corresponding author*

Key words: *differential evolution, optimization*

Motivation and Aim: Currently, the design of efficient algorithms and systems to solve the inverse problem of mathematical modeling continues to be a challenge due to large volume and heterogeneity of biomedical data, as well as high computational complexity of biomedical applications. It is well established, that an efficient optimization method should not only be fast and scalable across modern high performance architectures, but also reliable and robust.

Methods and Algorithms: ImDEEP is an improved Differential Evolution Entirely Parallel (DEEP) method developed in the work [1]. The Differential Evolution, introduced in 1995 by Storn and Price, considers the population, that is divided into branches, one per computational node [2]. The nodes are organized in a ring. The Differential Evolution Entirely Parallel method takes into account the individual age, that is defined as the number of iterations the individual survived without changes. We introduced the following improvements: (I) Allow several oldest individuals at (k + 1)th branch to be overwritten by the same number of best ones from kth branch. Communication period Π is a number of iterations between migrations. (II) In order to increase the robustness of the procedure we have implemented a new selection rule for Differential Evolution that allows us to use several different objective functions in offspring evaluation. The offspring replaces its parent if the value of the quality functional for the offspring set of parameters is less than that for the parental one. The additional objective functions are checked in the opposite case. The offspring replaces its parent if the value of some objective function is better and the randomly selected value is less than the predefined parameter for this function.

Results: We compared the performance of ImDEEP with original method and the state of the art optimization techniques such as Evolutionary Strategy (ES). The numerical results shows clear that ImDEEP outperforms the predecessors and is at least 2 times faster than ES.

Availability: on request from the authors.

Acknowledgements: We are very thankful to V. Gursky for many valuable discussions. The support of the study by the State Contract № 14.740.11.0166, RFBR Grants 11-04-01162, 10-01-00627, 11-01-00573, EU-FP7 SYSPATHO №260429 is gratefully acknowledged.

References:

1. Konstantin Kozlov, Alexander Samsonov (2011). DEEP – Differential Evolution Entirely Parallel Method for Gene Regulatory Networks, Journal of Supercomputing, Volume 57, pp.172-178.
2. Storn R., Price K., (1995), Differential Evolution – A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces, Technical Report TR-95-012, ICSI.

PROCESSING OF BIOMEDICAL IMAGES IN TeraPro

Kozlov K.N.*[1, 2], Baumann P.[3, 4], Waldmann J.[3], Samsonova M.G.[1, 2]
[1] *St.Petersburg State Polytechnical University, St.Petersburg, Russia;*
[2] *ProStack LLC, St.Petersburg, Russia;*
[3] *Jacobs University, Bremen 28759, Germany;*
[4] *Rasdaman GmbH, Bremen 28757, Germany*
e-mail: kozlov@spbcas.ru
**Corresponding author*

Key words: *processing and analysis of biomedical images*

Motivation and Aim: Image processing plays important role in medical diagnostics and prognosis of clinical outcomes, tissue and organ examination, measurement of therapeutic response, in clinical trails of new drugs, as well as in biology. Imaging systems today offer powerful processing technology, but typically are limited to image sizes of main memory. Picture Archiving and Communication Systems (PACS) are widely adopted for storage, retrieval, management, distribution, and presentation of electronic medical images. A common performance bottleneck in PACS currently is that the high-volume imagery is served through a low-intelligence server, with advanced processing concentrated on the client. This typically leads to excessive data transfer and reduced performance. *Methods and Algorithms:* In the TeraPro system we aim to combine the processing power of current imaging systems with the scalability and efficiency of current database technology and the user orientation of ontology-based user interfaces [1]. Integration of content-aware storage system with scalable and extensible image processing platform stretches the limits of biomedical image-based applications. The image processing package ProStack developed at St. Petersburg State Polytechnical University is capable for processing of 2D and 3D images. The extensible method repository includes a set of modules for noise reduction, edge detection, object recognition and classification, etc. Modules can be combined graphically into complex image processing workflows, saved, and reused. The system architecture allows integration with open source and commercial packages. The database is the rasdaman ("raster data manager") system [2], which is middleware (but often called a "database system" itself) storing any-dimensional raster data of unlimited sizes in a standard database and giving flexible access through a query language which extends standard SQL with declarative, optimizable raster expressions. *Results:* The system was successfully used to store and visualize images of lymphoma obtained with slide scanner. The uncompressed size of one image is about 50 Gb. The images can be viewed in the web interface of TeraPro on any internet device such as smartphone, tablet or netbook, without installation of the special software. *Availability:* http://urchin.spbcas.ru/trac/TeraPro *Acknowledgements:* We are very thankful to A. Pisarev and E. Myasnikova for many valuable discussions. The support of the study by the State Contract with FASIE № 14069, RFBR Grants 11-04-01162, 10-01-00627 is gratefully acknowledged.

References:

1. K.Kozlov, A.Pisarev, P. Baumann, M.Samsonova (2011). TeraPro, a System for Processing of Large-Scale Biomedical Images, Proc. of OGRW-8-2011, pp. 150-153
2. Baumann P (1994). On the Management of Multidimensional Discrete Data, VLDB Journal, №4, pp. 401-444.

MODELING OF GAP GENE EXPRESSION IN *DROSOPHILA KRUPPEL* MUTANTS

Kozlov K.N., Surkova S.Yu., Samsonova M.G.*
St. Petersburg State Polytechnical University, Russia
e-mail: samson@spbcas.ru
** Corresponding author*

Key words: *Drosophila, segmentation genes, mathematical modeling*

Motivation and Aim: The segmentation gene network in *Drosophila* embryo solves the fundamental problem of embryonic patterning: how to establish a periodic pattern of gene expression, which determines both the positions and the identities of body segments. The gap gene network constitutes the first zygotic regulatory tier in this process. Here we have applied the systems-level approach to investigate the regulatory effect of gap gene *Kruppel (Kr)* on segmentation gene expression.

Methods and Algorithms: We acquired a large dataset on the expression of gap genes in *Kr* null mutants and demonstrated that the expression levels of these genes are significantly reduced in the second half of cycle 14A. To explain this novel biological result we applied the gene circuit method which extracts regulatory information from spatial gene expression data.

Results: Previous attempts to use this formalism to correctly and quantitatively reproduce gap gene expression in mutants for a trunk gap gene failed, therefore here we constructed a revised model and showed that it correctly reproduces the expression patterns of gap genes in *Kr* null mutants. We found that the remarkable alteration of gap gene expression patterns in *Kr* mutants can be explained by the dynamic decrease of activating effect of Cad on a target gene and exclusion of *Kr* gene from the complex network of gap gene interactions, that makes it possible for other interactions, in particular, between *hb* and *gt,* to come into effect.

Conclusion: The successful modeling of the quantitative aspects of gap gene expression in mutant for the trunk gap gene *Kr* is a significant achievement of this work. This result also clearly indicates that the oversimplified representation of transcriptional regulation in the previous models is one of the reasons for unsuccessful attempts of mutant simulations.

Availability: All data are available from authors.

SYSMO-DB: A COMMUNITY-BASED APPROACH TO DATA SHARING

Krebs O.*[1], Wolstencroft K.[2], Owen S.[2], Nguyen Q.[1], du Preez F.[2], Mueller W.[1], Goble C.[2], Snoep J.L.[2]

[1] *Heidelberg Institute for Theoretical Studies (HITS), Germany;*
[2] *University of Manchester, United Kingdom;*
e-mail: olga.krebs@h-its.org
** Corresponding author*

Key words: *data management, data integration, standardisation, ontology, systems biology*

Motivation and Aim: Systems biology research is typically performed by multidisciplinary groups of scientists, often in large consortia and in distributed locations. There is a growing requirement for exchanging experimental data, mathematical models, and scientific protocols between consortium members and a necessity to record and share the outcomes of experiments and the links between data and models. The overall output of a research consortium is also a valuable commodity in its own right..

Results: The SEEK is an open-source, web-based platform for the management and exchange of Systems Biology data, models and processes. It was originally developed in the SysMO-DB project (http://www.sysmodb.org) for the pan-European SysMO consortia (Systems Biology of Micro Organisms). However, it is now also being adopted by a large number of other consortia across Europe, for example, the Virtual Liver, EviMalar and Unicellsys.

The SysMO-DB solution is being developed in close collaboration with the users, and with a very pragmatic approach, trying to adapt to the common procedures of users and combining this with methods and technologies that allow an effective dissemination, linkage and exchange of data and information.

Underlying the SEEK is the JERM (Just Enough Results Model), which is a minimum information model describing the the structure and content of the SEEK assets and relationships between them. A JERM for any one type of data (i.e. microarray data, or metabolomic data) is the minimum data schema that SysMO projects agree to share. This is used to create JERM templates. SysMO-DB leverages these minimum models wherever possible, enabling easy export and publishing of SysMO data to public repositories. In addition, we have developed RightField (the Spreadsheet Ontology Annotation tool) to help researchers to develop templates for experimental data with embedded semantic descriptions. Data can then be consistently annotated with terms from appropriate ontologies and controlled vocabularies consistently.

The SEEK provides an access control layer to enable researchers to restrict access to their data to collaborators and colleagues or to share it with the wider community.

Availability: SEEK code is open source and available for download http://code.google.com/p/sysmo-db/. For a demonstration of the SEEK capabilities, and to try out the software, demo SEEK is available here https://demo.sysmo-db.org/. RightField download: http://www.sysmo-db.org/rightfield

References:

1. Wolstencroft K, Owen S, du Preez F, Krebs O, Mueller W, Goble C, Snoep JL..Uthor, O.N.E.More. (2011) The SEEK: a platform for sharing data and models in systems biology., In: Methods Enzymol.; v.500, pp. 629-55.

ELECTROSTATICS AND BENDING IN PROMOTER FUNCTIONING DURING THE GLOBAL METABOLIC SWITCH

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, DNA electrostatic properties, genome physical properties, promoters*

Motivation and Aim: One of the studied parameters that influence promoter strength is the DNA bending in promoter area. *E. coli* BNT2 promoter relies on the DNA bending in its functioning. This promoter has an A-tract in the spacer and unusual AAAAAT -10 element. In [1] substituted the spacer AAA tract with the TCG and found that the resultant promoter posses a) lesser bending and b) much lower activity. As the activity of native bent promoter strongly reduces while the temperature rises above the curvature-relaxing point, authors conclude the promoter activity relies on the intrinsically bent DNA structure of AT tracts. They confirmed this by a) calculating the curvature and b) by studying the DNA retardation in electrophoresis under different temperatures. However, the twice lower activity of mutant promoter under the temperature above the curvature-relaxing point, which can not be accounted to the bending, was unexpected for the authors and remained unexplained.

Methods and Algorithms: DEPPDB and its tools [2,3] were used to carry out the analysis.

Results: The native promoter with AT tract posses a prominent rise in electronegative potential value along it, that correlates with the promoter strength, probably facilitating the promoter recognition and binding. In the mutant with the part of the A tract substituted with the TCG sequence there is no such a rise in this place, that obviously affects the DNA-RNA polymerase interactions. Thus the electrostatic properties of promoter DNA complement the DNA bending in E. coli O157:H7 pO157 plasmid BNT2 promoter functioning.

Conclusion: The given example of physical properties cooperation is of particular interest to the studies of well-known phenomenon of global metabolic switch during the inside (host) to outside (environment) transition of some symbiotic and pathogenic bacteria. Any other transitions and global regulations involving differences in temperature could also be liable to these effects. This could be studied by means of DEPPDB and its tools.

Acknowledgements: The authors are grateful to IMPB RAS for hosting the Database.

References:

1. J. W. Yoon, M. K. Park, C. J. Hovde, Seung-Hak Cho, Jong-Chul Kim, Mi-Sun Park, W. Kim, ”Characterization of BNT2, an intrinsically curved DNA of Escherichia coli O157:H7,” Biochem. Biophys. Res. Commun. 391(4), 1792-7 (2010).
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

ELECTROSTATIC AND OTHER PHYSICAL PROPERTIES OF TRANSCRIPTION FACTORS 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, DNA electrostatic properties, genome physical properties, promoters*

Motivation and Aim: Electrostatic and other physical properties of genome DNA are well recognized to influence its interactions with different proteins, in particular regarding of transcription regulation. To reveal the role of these properties in the transcription regulation proteins binding we studied binding sites of different families of transcription factors in different prokariotic taxa.

Methods and Algorithms: DEPPDB – DNA Electrostatic and other Physical Properties Database and its tools [1,2] were used to carry out the analysis.

Results: The averaged profiles of the DNA electrostatic 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.

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: These observations support the idea of a significant and universal role of electrostatics in the regulations of cell genome functioning and evolution, and demonstrate the utility of DEPPDB to study these phenomena. It is tempting to hypothesize that this may facilitate the process of horizontal gene transfer and adaptation of new regulatory circuits and thus being important 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

NEW EVIDENCES OF THE ELECTROSTATIC NATURE OF PROMOTERS UP-ELEMENT COMBINED WITH ITS OTHER PHYSICAL PROPERTIES

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, DNA electrostatic properties, genome physical properties, promoters*

Motivation and Aim: One of the elements that may play a crucial role in the promoter strength regulation is a so-called "up-element", which interacts with the alpha-subunit of RNAP and thus facilitates its binding to the promoter. There is no text consensus in the "up-element" (though high AT content is often attributed) and functionality of this region is defined by its physical properties. We have shown earlier, that electrostatics is responsible for its functioning during the global transcription switch under the T4 infection and that strong T4, early T7-like and *E.coli* ribosomal promoters with pronounced up-element have high levels of the electrostatic potential within it. To further investigate the electrostatic nature of the up-element and reveal the role of other physical properties in its functioning we analyzed promoters of different strength of phage lambda and its relatives and a series of promoters under mutagenesis taken from literature.

Methods and Algorithms: DEPPDB – DNA Electrostatic and other Physical Properties Database and its tools [1,2] were used to carry out the analysis.

Results: 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. Most interesting is the example of lambda-like phages strong pL promoters. They all possess strong electrostatic up-elements, the sequence texts of which are quite different.

Strong promoters such as rrnB with eliminated up-element (and thus greatly reduced strength) do not have pronounced electrostatic valleys in the corresponding area. Mutated up-elements with enhanced promoter strength exhibit deep electrostatic valleys and peculiarities of some other physical properties, though the dependence of promoter strength on the electrostatic up-element intensity is not linear in very strong promoters. This may indicate the complex nature of the promoter functioning.

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

COMPLEX GENOME SEQUENCING: PRELIMINARY DATA OF SIBERIAN LARCH COMPLETE GENOME SEQUENCING

Krutovsky K.V.*[1, 2, 5], Vaganov E.A.[2], Chubugina I.V. [2, 3], Oreshkova N.V.[4], Tretyakova I.N.[4], Tyazhelova T.V. [5]

[1] *Department of Ecosystem Science and Management, Texas A&M University, College Station, TX, USA;*
[2] *Genome Research Center, Siberian Federal University, Krasnoyarsk, Russia;*
[3] *Center for Forest Protection, Krasnoyarsk, Russia;*
[4] *V.N. Sukachev Institute of Forest, SB RAS, Krasnoyarsk, Russia;*
[5] *N.I. Vavilov Institute of General Genetics, RAS, Moscow, Russia*
e-mail: k-krutovsky@tamu.edu
** Corresponding author*

Key words: *de novo sequencing, complex genomes, conifers*, *Larix sibirica Ledeb*

Motivation and Aim: The purpose of the study is to completely sequence, assemble and annotate *de novo* the Siberian larch (*Larix sibirica* Ledeb.) genome (1C = 12.03 Gbp), which is four times larger than human genome (1C = 3.20 Gbp) that remains the largest genome completely sequenced so far. Meanwhile, larch is one of the most important key elements of Siberian boreal forests that have great economic and ecological values. However, the study of larch and other closely related important conifer forest tree species is hindered by almost complete lack of data on its genome structure and genes that control important adaptive and selective traits. Complete larch genome sequence would allow us to obtain such data and effectively use them for studying conifer forests genetic variation, genetic adaptation to global climate change and for creating conservation and breeding programs.

Methods and Algorithms: The gigantic genome size of conifers and high allelic variation impede their complete genome sequencing and assembling. The conifer genomes are not only extremely large, but also contain a great number of repetitive elements and large gene families with high similarity in nucleotide sequences. To overcome these problems and facilitate assembling we use an innovative unique approach via using haploid tissue cultures developed from haploid immature megagametophytes (female gametophytic tissue).

Results: The haploid nature of tissue cultures or calluses obtained from megagametophytes was confirmed by genotyping their nuclear genomic DNA with informative SSR markers that are heterozygous in the diploid tissue of the parent tree. After fragmentation the fraction of nuclear genomic DNA within 550-600 nucleotide base pairs size range was used for paired-end sequencing with 101 cycles and four lanes of a flow cell of the *Illumina HiSeq 2000* sequencer that should give an expected ~12X genome coverage. The preliminary results based on these sequence data will be presented.

Conclusion: The obtained data represent the first step in the multi-disciplinary integrative innovative international project on complete *de novo* larch genome sequencing that is planned to be done in the Genome Research Center recently established at the Siberian Federal University (Krasnoyarsk, Russia).

Acknowledgements: This work was supported by grant from Russian government department of Science and Education to Siberian Federal University «The genetic researches of the Siberian larch».

EVOLUTION OF EXON-INTRON GENE STRUCTURE AND ALTERNATIVE SPLICING: WHAT WE CAN LEARN FROM COMPLETELY SEQUENCED GENOMES AND PREDICT FOR NON-MODEL SPECIES

Krutovsky K.V.*[1, 2, 3], Koralewski T.E. [1]
[1] *Texas A&M University, College Station, TX, USA;*
[2] *Genome Research Center, Siberian Federal University, Krasnoyarsk, Russia;*
[3] *N.I. Vavilov Institute of General Genetics, RAS, Moscow, Russia*
e-mail: k-krutovsky@tamu.edu
** Corresponding author*

Key words: *evolution, exon-intron gene structure, alternative splicing, proteomic and metabolomic complexity, comparative genomics, regression models*

Motivation and Aim: Diversity, complexity and difference in anatomy, physiology and behavior are greatly increasing from relatively simple, less evolved organisms toward more evolutionary advanced species. However, this does not affect significantly the number of genes that demonstrate a moderate increase in genomes of completely sequenced model organisms. For instance, *Arabidopsis* and human have a similar number of genes. This suggests that regulatory and other post-transcriptional processes might play an increasingly more important role in maintaining complexity of more evolved organisms. Alternative splicing is one of the most important post-transcriptional processes that increases proteomic and metabolomic complexity without increasing number of genes. Therefore, our main objective was to test the hypothesis that AS contributes greatly to proteomic and metabolomic complexity of more evolutionary advanced organisms. One of our objectives was also to estimate the AS rate, exon-intron size and their numbers across evolutionary distant species and to use these parameters for developing a computer model that would allow us to predict metabolomic and proteomic complexity in the non-model organisms.

Methods and Algorithms: To test our hypothesis we developed a computer program that allowed us to infer the types and frequency of AS in evolutionary different completely sequenced and fully annotated species from the NCBI GenBank files.

Results: We performed a genome-wide comparative computer analysis of AS types, number of genes, gene products and exons in 36 completely sequenced model species. We created statistical regression models to fit these data and applied them to non-model species whose genomes have not been completely sequenced yet (Koralewski & Krutovsky 2011).

Conclusion: Number of exons, AS rate and diversity and, therefore, proteomic complexity are significantly increased in more evolutionary advanced complex organisms. Our data demonstrate a great diversity of AS types. The most common 10 types represent only ~12% of the total number of AS events. The most common AS types are due to alternative 3' splice site and alternative promoters. Using our models, the genome-wide characteristics, such as exon length and exon-gene ratio, can be predicted based on parameters estimated from available genomic data.

Availability: Perl scripts specifically written for this study are available at http://treenome.tamu.edu.

References:

1. T.E. Koralewski, K.V. Krutovsky (2011) Evolution of Exon-Intron Structure and Alternative Splicing. *PLoS ONE* **6**(3): e18055. doi:10.1371/journal.pone.0018055

IS THE ASSOCIATION BETWEEN -308G->A TNF-α AND MULTIPLE SCLEROSIS INDEPENDENT OF HLA-DRB1*15?

Kudryavtseva E.A.*, Rozhdestvenskii A.S., Kakulya A.V., Khanokh E.V., Malkova N.A., Korobko D.S., Platonov F.A., Aref'eva E.G., Zagorskaya N.N., Alifirova V.M., Titova M.A., Smagina I.V., El′chaninova S.A., Zolovkina A.G., Puzyrev V.P., Tsareva E.Y., Favorova O.O., Boiko A.N., Filipenko M.L.

Institute of Chemical Biology and Fundamental Medicine SB RAS, Novosibisk;
Novosibirsk State University, Novosibirsk; Omsk State Medical Academy, Omsk;
Novosibirsk Oblast State Clinical Hospital, Novosibirsk;
Republican Hospital No. 2, Ministry of Health, Sakha Republic, Yakutsk;
State Health Facility Kemerovo Oblast Clinical Hospital, Kemerovo;
Siberian State Medical University, Tomsk;
Research Institute of Medical Genetics, SD, RAMS, Tomsk;
Altai State Medical University, Barnaul; Moscow center of multiple sclerosis on the basis of hospital № 11
e-mail: kudryavtseva_ekaterina@ngs.ru

** Corresponding author*

Key words: *multiple sclerosis, TNFa, HLA-DRB1*

Motivation and Aim The tumor necrosis factor α (TNF-α) is known as proinflammatory cytokine implicated in the pathogenesis of autoimmune and infectious diseases. The gene of TNF-α are located in the region HLA class III. TheTNF-α promoter polymorphism-308 G->A in the human has been most often used in the associated studies. In view of gene location, it has been speculated that polymorphism of this locus might contribute to HLA association with multiple sclerosis (MS). The aim of this study is an assessment of chosen factors of genetic susceptibility to MS such as HLA-DRB1 and TNF-α polymorphism-308(G/A) in residents RF. More over we would like to estimate whether possible relationship of TNF-α-308(G/A) is due to a primary association or mediated by LD with the susceptible DRB1 alleles.

Methods and Algorithms A group of 1650 MS patients and 992 healthy peoples were randomly selected as the subject of this study. Determination of genotypes of TNF-α-308(G/A) was performed by Real-time PCR. Determination of genotypes of HLA-DRB1 was performed by sequencing. Statistical analysis was performed using the R-language.

Results The G allele of -308G->A TNF-α is associated with MS (OR=1.34 [1.10-1.64] p= 0.004). A complete meta-analysis of all analogous studies published to date showed that the G allele is a risk (OR=1.29, 95%C.I.=[1.13-1.46], $p=9*10^{-5}$). The HLA-DRB1*15 is associated with MS (OR=2.40, 95%C.I.=[1.60-3.61], $p=3*10^{-5}$). The -308G->A TNFa and HLA-DRB1*15 are in linkage disequilibrium (D'=.78, $r^2=0.02$, $\chi^2=5.53$). Using LRT we have shown the HLA-DRB1*15 is key risk factor between -308G->A TNFa and HLA-DRB1*15.

Conclusion Association -308G->A TNF-α and MS is induced by *HLA-DRB1*15*.

COMPREHENSIVE COLLECTION OF HUMAN TRANSCRIPTION FACTOR BINDING SITE MODELS

Kulakovskiy I.V.*[1, 2], Medvedeva Y.A.[3], Kasianov A.S.[1, 2], Vorontsov I.E. [4, 1], Schaefer U.[3], Bajic V.B.[3], Makeev V.J. [1, 2]

[1] *Vavilov Institute of General Genetics RAS, Moscow, Russia;*
[2] *Engelhardt Institute of Molecular Biology RAS, Moscow, Russia;*
[3] *King Abdullah University of Science and Technology, Thuwal, Jeddah, Saudi Arabia;*
[4] *Moscow Institute of Physics and Technology, Moscow, Russia*
e-mail: ivan.kulakovskiy@gmail.com
** Corresponding author*

Key words: *transcription factor binding sites, transcription factor binding models, TFBS model collection, motif discovery, position weight matrices, ChIP-Seq*

Motivation and Aim: Existing collections of transcription factor binding site (TFBS) models often contain multiple models for the same transcription factor (TF). Models obtained by different experimental techniques can have peculiar biases related to the properties of experimental methods. Such biases may be compensated to an extent by data integration and simultaneous usage of binding site sequences from various experiments. At the same time having a single binding site model for each TF is very desirable for further applications.

Methods and Algorithms: DNA sequences of TF binding regions obtained by both pregenomic and high-throughput methods were collected from existing databases (such as TRANSFAC and JASPAR) and other public data. ChIPMunk software [http://autosome.ru/ChIPMunk] was used to construct positional weight matrices. Four motif discovery strategies were tested based on different motif shape priors including flat and periodic priors associated with DNA helix pitch. A quality rating was manually assigned to each binding site model based on known binding preferences. An appropriate model was selected for each TF, with similar models selected for related TFs. In any case only one model per transcription factor was selected unless there was additional evidence for two distinct binding models or different stable modes of dimerization.

Results: HOmo sapiens COmprehensive MOdel COllection (HOCOMOCO) of transcription factor binding models is constructed by careful integration of data from different sources. HOCOMOCO contains curated non-redundant positional weight matrices for four hundreds of human TFs where more than 150 models are based on more than one data source.

Conclusion: We provide HOCOMOCO collection with UniPROT links for all binding models and initial binding segments used to construct positional weight matrices. This should facilitate usage of the HOCOMOCO collection in further automated analyses.

Availability: http://autosome.ru/HOCOMOCO

Acknowledgements: We personally thank Ivan Lysov (Trafica, LLC) and Prof. A. S. Kondrashov (Faculty of Bioengineering and Bioinformatics, M.V. Lomonosov Moscow State University) for providing computational facilities.

Funding: This work was supported, in part, by a Dynasty Foundation Fellowship to I.V.K.; Presidium of the Russian Academy of Sciences program in Cellular and Molecular Biology; Russian Ministry of Science and Education State Contracts (07.514.11.4005 and 07.514.11.4006); Russian Ministry of Science and Education grant (11.G34.31.0008).

WHOLE GENOME SEQUENCING AND PHYLOGENETIC ANALYSIS OF *VIBRIO CHOLERAE* O1 ELTOR INABA № 301 STRAIN

Kuleshov K.V.*[1], Shipulin G.A.[1], Markelov M.L.[1], Dedkov V.G.[1], Podkolzin A.T.[1], Vodop'ianov S.O.[2], Kermanov A.V.[2], Kruglikov V.D.[2], Mazrukho A.B.[2], Vodop'ianov A.S.[2], Pisanov R.V.[2]

[1] Federal Budget Institute of Science "Central Research Institute for Epidemiology", Moscow, Russia;

[2] The Federal Government Health Institution "Rostov-on-Don Plague Control Research Institute", Rostov-on-Don, Russia

e-mail: konstantinkul@gmail.com

** Corresponding author*

Key words: *whole genome sequencing, Vibrio cholerae, core genome phylogeny*

Aim: The aim of this research was to determine molecular genetic markers for the understanding of the possible origin of Vibrio cholerae O1 Eltor Inaba № 301 (#2011EL-301) strain isolated from sea water in the city of Taganrog in the summer of 2011.

Methods and Algorithms: Whole genomic sequencing protocol included the following steps: sequencing of fragment libraries using both Roche 454 GS Junior(Roche Diagnostics) and MiSeq (Illumina) sequencers; *de novo* assembling contigs from Roche's single reads using Newbler 2.5; mapping the Illumina pair-end reads to Roche's assembled contigs with CLC Genomics Workbench v5.5.

Results: Using Roche 454 reads 124 contigs were assembled *de novo* the N50 was 135 kb, average coverage was 33x. For validation and correction of significant SNPs on assembled contigs we mapped Illumina pair-end reads to this contigs. The consensus sequence was used for further analysis. The average coverage for 124 contigs was 100x. The isolate under analysis contains a hybrid prophage CTX localized in chromosome I. This prophage carries the classical type allele of ctxB and rstR El-Tor allele. VPI-1 region carries tcpA El-Tor allele. Loci (VSPI, VPI-1, VPI-2, SXT) have high similarity level with recently isolated strains (1). Comparison of assembled contigs with several complete and incomplete genomes of different *V. cholerae* strains from the GenBank database using Progressive Mauve algorithm revealed the most similar genomes which were isolated from Haiti, Africa and Asia (1). In compliance with earlier proposed scheme for SNP typing of Vibrio cholerae strains of 7th pandemic, which is based on 30 SNPs, our strain belongs to group V. 29 genomes were analyzed on the basis of common ortholog genes to understand in more detail the phylogenetic relations of 2011EL-301 strain among recent isolates of V. cholerae. As a result of the core genome phylogeny our isolate was included in one cluster with strain associated with cholera cases from South Africa in 2009 (# 2011EL-1137) and cases of transmission of cholera from Pakistan (#3582-05, #2009V-1116, #2009V-1046, #2010V-1014). In comparison with CRIS101 genome sequence in core genome phylogenetic tree and data from earlier published study (3) our strain refers to the 3th wave of the 7th pandemic. Whole Genome Shotgun project has been deposited at DDBJ/EMBL/GenBank under the accession AJFN00000000.

References:

1. A.R. Reimer et al. (2011), *Emerg Infect Dis.*,**17(11)**:2113-21
2. C. Lam et al. (2011) *Emerg Infect Dis.,***16(7)**:1130-2
3. A. Mutreja et al. (2011), *Nature,* **477**, 462–465

ASSOCIATION OF *ITGB3* AND *GNB3* VARIANTS WITH THE DEVELOPMENT OF VASCULAR COMPLICATIONS IN PATIENTS WITH ACUTE CORONARY SYNDROME

Kulish E.V.*[1], Makeeva O.A.[1, 2], Golubenko M.V.[1, 2], Zykov M.V. [2], Kashtalap V.V.[2]
[1] *Research Institute of Medical Genetics, SB RAMS, Tomsk, Russia;*
[2] *Research Institute of Complex Problems of Cardiovascular Diseases, SB RAMS, Kemerovo, Russia*
e-mail: elena-kulish@medgenetics.ru
** Corresponding author*

Key words: *acute coronary syndrome, single nucleotide polymorphism, arrayed primer extension-based genotyping method, DNA microarray*

Motivation and Aim: Acute coronary syndrome is a common disorder and is a significant cause of morbidity and mortality worldwide.

The aim of the study was to examine association of genetic polymorphisms in *ACE, AGT, AGTR1, CEPT, EDN1, F2, F5, GNB3, ITGB3, LIPC, LPL, MTHFR, NOS3, PON1, PPARG, TCF7L2, MTND2, TNF, CDKN2A/B* with cardiovascular complications in patients with acute coronary syndrome (n=171) in one year observational period after acute event.

Methods and Algorithms: The choice of SNPs was based on results of association studies, analysis of electronic databases, most of markers were confirmed on local population. Arrayed primer extension based genotyping with Genorama™ Imaging System (Asper Biotech Ltd) was used for SNP detection.

Results: A comparative analysis of allele frequencies in patients with acute coronary syndrome with and without vascular complications showed statistically significant higher *ITGB3* gene (rs5918) allele C frequency (p=0.044). Differences in genotype distribution of *ITGB3* (rs5918) and *GNB3* (rs5443)had also been detected in patients with and without vascular complications(p=0,007 and p=0,042, respectively).

Conclusion: ITGB3 (rs5918) and *GNB3* (rs5443) may play a role in development of vascular complications in patients with acute coronary syndrome.

GENETIC DIVERSITY IN EXTREMOPHILIC BACTERIAL COMMUNITY FROM HOT SPRING «URITSKY», BAIKAL

Kurilshikov A.M. [1], Babkin I.V. [1], Morozova V.V. [1], Bryanskaya A.V. [2], Tikunov A.Yu.[1], Lazareva E.V. [3], Zhmodik S.M. [3], Tikunova N.V. [1]

[1] Institute of Chemical Biology and Fundamental Medicine SB RAS, Novosibirsk,Russia;
[2] Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia;
[3] Institute of Geology and Mineralogy SB RAS, Novosibirsk, Russia
e-mail: alexa.kur@gmail.com

Key words: *metagenomics, extremophile, bacteria, 16s rRNA, Baikal*

Environmental sample for this study was obtained from Uritsky hot spring located in the Barguzin basin of the Baikal rift zone. DNA was isolated and then amplified with primers specific to 16s ribosomal RNA gene fragment with 1230 nucleotides of estimated length. Purified amplicons were cloned into E.coli plasmids. Since monoclonal libraries were grown, inserted fragments were sequenced by Sanger.

In total, 242 clones were sequenced. All sequenced clones were chimera checked by Mallard (Cardiff University, Great Britain) and Bellerophon v.3 (GreenGenes project) software packages. 14 sequences were identified as chimeras. To calculate richness of the studied environmental sample, rarefaction analysis was performed with MOTHUR package. Number of OTU (with 95% cutoff) in sample was evaluated as 180 by Chao estimator. Phylogenetic analysis was performed by RDP software (http://rdp.cme.msu.edu) with 80% confidence threshold. 47 sequences were indetified as *Deinococcus*, 20 – *Cyanobacteria* , 25 – *Bacteroidetes*, 77 – *Proteobacteria*, 17 – *Actinobacteria*, 30 – *Planctomycetes*, 16 – *Chloroflexi*, 2 – *Fusobacteria*, 1 – OD1, 6 – *Firmicutes*. In addition, there were several *Proteobacteria* phylum members: *Alpha* – 21, *Beta* – 19, *Gamma* – 30 and 2 *Deltaproteobacteria*. 46 sequences couldn't be determined on family level; 5 *Proteo-* and 5 *Eubacteria* couldn't be determined to any class. The greatest difference between sequenced bacteria and reference sequences from GenBank (NCBI), RDP and other databases was revealed in *Cyanobacteria* phylum. None of the sequenced *Cyanobacteria* can be related to any family. Sequences with highest homology to these *Cyanobacteria* were discovered in hot springs from different places: Yellowstone (USA), Kumamoto (Japan), Rincon de la Vieja (Costa-Rica).

CHANGES IN PROTEIN COMPOSITION OF HUMAN URINE AFTER PROLONGED ORBITAL FLIGHTS

Larina I.M.*[1], Nikolaev E.N.[2], Pastushkova L.H.[1], Valeeva O.A.[1],
Kononihin A.S.[2], Kireev K.S.[3], Tiys E.S.[4], Ivanisenko V.A.[4], Kolchanov N.A. [4]
[1] Institute of Medicobiologic Problems Russian Federation State Scientific Research Center RAS, Moscow, Russia;
[2] Emanuel Institute of Biochemical Physics RAS, Moscow, Russia;
[3] Gagarin Center of Cosmonauts Training, Star City, Russia;
[4]I nstitute of Cytology and Genetics SB RAS, Novosibirsk, Russia
e-mail: irina.larina@gmail.com
** Corresponding author*

Motivation and Aim: Changes in the human body during space flights affect all physiological systems, including changing the protein composition of body fluids.

Methods and Algorithms: To investigate the adaptive plasticity of the protein composition of urine we analyzed its samples obtained from six Russian cosmonauts at the age of 35 to 51 years who have completed space missions lasting from 169 to 199 days on the International Space Station (ISS). The analysis of mixtures peptides derived from the samples was carried out by chromato-mass spectrometric method, using the accurate mass and time tag retention of peptides in the chromatographic column (TMMVU). Of the total number of detected peptides (430) twenty-one peptide was presented in the samples of the cosmonauts urine at all stages of the survey. After the establishment of their matching proteins (as UniProt), an analysis of their cellular localization, tissue specificity and functions was carried out.

Results: It was found that the presence of the peptides in the groups of samples (preflight, post-flight) is different, and thus, it was detected a certain drift of the protein-peptide composition of urine, caused by prolonged space flight. 209 peptides, based on a database Tissue-specific Gene Expression and Regulation (TiGER), were characterized as belonging to different tissues. Proteins of liver tissue, bone and soft tissues were significantly more represented in cosmonauts urine in comparison with those in the database TiGER ($P<0,05$) according to Bonferroni adjusted Fisher's exact test for multiple comparisons. Also on the basis of analysis of composition of proteins specifically expressed in the kidney, the Gene Ontology processes were identified that are specific to the background period, for the first and seventh days after the flight.

Conclusion: Proteomic study of the composition of urine, performed by a highly sensitive proteomics methods have provided new data required to clarify the origin of changes in the human body, occurring under the influence of space flight.

EFFECTS OF IN- AND OUTBREEDING IN POPULATIONS OF DIPLOID ORGANISMS: COMPUTER SIMULATIONS WITH THE DIPLOID EVOLUTIONARY CONSTRUCTOR

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: *computer simulations, modeling, inbreeding, outbreeding, evolution*

Motivation and Aim: A large part of the genetic component of the phenotypic variance is the result of coordinated functioning of genes, which form gene networks and particularly regulatory contours of gene networks. Allelic combinations optimizing certain functions in gene networks are selected in the course of evolution. However some population processes such as outbreeding may break stable allelic combinations, which in its turn leads towards the increase of the reaction norm [1] and, in some cases, decrease of progeny fitness. Similar decrease of the fitness can also be observed in the case of inbreeding. This study was aimed to construct and analyze the computer models of the inbreeding and outbreeding processes within populations of diploid organisms with due account for various types of heredity.

Methods and Algorithms: The software package "Diploid evolutionary constructor" (DEC) was used to perform computer simulations.

Results: We have developed and studied computer model of the inbreeding and outbreeding processes within a population of diploid model organisms. The haploid genome of the organism contained six genes: *A, B, C, D, E, F*. The organism's phenotype was characterized by two complex traits – fertility and disease resistance. Simple quantitative traits A and B were coded by genes *A* and *B*, respectively, and the fertility itself was defined according to the cumulative effect of those traits. Disease resistance was defined by four simple quantitative traits C, D, E, F, which coded by the corresponding genes. All simple traits were calculated using various types of dominance: full dominance, codominance, maternal-effect dominance. Additive and multiplicative effects were also considered.

Conclusion: It has been shown that co-adaptive genes may evolve under the influence of both inbreeding and outbreeding. The pressure of inbreeding within one group of genes (for instance, included in the same gene network) may affect the evolution end genetic diversity of another genes including co-adaptive ones (even if they are in other network(s)). It has also been shown that the evolution of co-adaptive genes (in diploid genome) is affected by the dominance type.

Acknowledgements: The work was supported by the RFBR grants 10-04-01310-a, 12-07-00671-a, RAS Program № 28.

References:

1. Yu.E. Dubrova, L.V. Bogatyryova (1993) Hybridity effects on variation of anthropometric traits in newborns, *Genetika,* **29:** 1702-1711.

WHEN GENE NETWORKS MAY NOT WORK: COMPUTER MODELING WITH THE HAPLOID EVOLUTIONARY CONSTRUCTOR

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: *gene network, evolution, modeling, computer simulations*

Motivation and Aim: Mathematical and computer modeling of gene networks functioning becomes more and more widely used tool in different branches of biology. Models constructed with the use of various approaches help us to obtain new knowledge about biological systems, to formulate new hypotheses, which often achieve experimental verification. Developed *in silico* at first and only then *in vitro* and *in vivo*, the genetic constructions show a good agreement between theory and experiment [1]. Meanwhile, it is obvious that such an ideal situation is rather exception than the rule. In this study we simulated and analyzed the evolution of a population of prokaryotic cells containing a certain genetic construction, and found limitations in which gene networks may not work.

Methods and Algorithms: The models of prokaryotic population of cells consuming on nonspecific substrate and producing two specific substrates were constructed using the software package "Haploid evolutionary constructor" – HEC (available at http://evol-constructor.bionet.nsc.ru). Synthesis of specific substrates in cells was described via gene network of molecular trigger [1].

Results and Conclusion: Simulations showed feasibility of appearance of trigger modes in the model depending on gene network parameters. Similar dependency was also shown on physiological, population and ecological parameters. The functioning regimes of gene networks within the organism were demonstrated to be significantly distinct from the regimes predicted on the basis of mathematical analysis of the corresponding gene networks models. In the molecular trigger model, the unstable steady state is of the saddle type. Under the HEC simulations we found the set of initial data for which the system stabilizes in this state. Such stabilization is possible due to additional factors. In particular, the range of cell wall permeability may constrain influence of substrates-trigger switches. Therefore, our models suggest an additional over-genetic mechanism for sustaining the stability of gene networks functioning regimes. The look-alike mechanisms were theoretically and experimentally investigated formerly [2]. On the one hand, they can explain "nonworking" of genetic constructions *in vivo*, while these constructions should work on the base of *in silico* calculations. On the other hand, these mechanisms are of great evolutionary importance and need further investigations.

Acknowledgements: RFBR grants 10-04-01310-a, 12-07-00671-a, RAS Program № 28.

References:

1. T.S. Gardner et al. (2000) Construction of a genetic toggle switch in *E.coli*, *Nature,* **403:** 339-342.
2. R.N. Tchuraev et al. (2000) Epigenes: design and construction of new heredity units, *FEBS Lett*, **486:** 200-202.

EVOLUTION IN PROKARYOTES-PHAGES COMMUNITIES: COMPUTER MODELING WITH THE HAPLOID EVOLUTIONARY CONSTRUCTOR

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: *prokaryotic community, phage, evolution, modeling, computer simulations*

Motivation and Aim: Prokaryotes and prokaryotic communities are suggested to be the most ancient form of the living matter on the Earth. According to the symbiogenetic theory of evolution, the complication of prokaryotes could be the result of several symbiogenetic acts. The leading role of those processes could be played by a horizontal gene transfer, which in its turn could be promoted by phages and other infection agents. In this study we performed the computer simulations of evolution of prokaryotic communities infected by phages, and then analyzed changes in evolutionary trends caused by infection.

Methods and Algorithms: The software package "Haploid evolutionary constructor" (HEC) was used to perform computer simulations.

Results and conclusion: In order to analyze the influence of phage infections on feasible evolutionary trends we have built several models of prokaryotic community infestation. As the base model we considered the model of trophic ring-like community consisted of three population, which fed each other. During the community evolution the processes of horizontal genes transfer and genes loss were stochastically modeled that led to the origin of novel populations. The addition of a phage population to the community caused the infection of all populations, while the proportion of infected cells depended on both the phage concentration and prokaryotic cells concentration in the medium. The probability that an infected population or its part would develop according to the lytic scenario depended on the conditions under which its parent population was at the moment of infection: if it steadily grew, then all infected cells in this subpopulation were lysed and released a new batch of phages; contrariwise, if the population was under the pessimal conditions and decreased in size, then all, or some of infected cells switched to the lysogenic cycle becoming on the one hand, carriers of the phage genes and on the other hand, immune to this type of phage. As a consequence of this, the infestation essentially changed evolutionary dynamics of the community by keeping down growth or even killing rapidly growing populations, hence preserving populations which are less competitive under current conditions. Therefore, phage infections played and play a key role in prokaryotic evolution.

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

Acknowledgements: The work was supported by the RFBR grants 10-04-01310-a, 12-07-00671-a, RAS Program № 28.

References:

1. S.A. Lashin et al. (2011) Trends in the Prokaryotic Community and Prokaryotic Community–Phage Systems, *Russian journal of genetics,* **47:** 1487-1495.

IN SILICO VERIFICATION OF CHIP-SEQ DATA

Levitsky V.G.*[1], Oshchepkov D.Y. [1], Vasiliev G.V. [1], Ershov N.I. [1], Merkulova T.I. [1], Kulakovskiy I.V.[2], Makeev V.J.[2]

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

[2] *Institute of Genetics and Selection of Industrial Microorganisms, Moscow, Russia*

e-mail: levitsky@bionet.nsc.ru

** Corresponding author*

Key words: *ChIP-Seq experiments, transcription factor binding sites recognition*

Motivation and Aim: ChIP-Seq data analysis requires application of recognition tools appropriate for transcription factor binding sites (TFBS) recognition. TF FoxA is critical to the development and function of the liver.

Methods and Results: Nucleotide sequences of 81 functional FoxA BS were retrieved from literature. The training sets of 53 BSs on the basis of degenerate motif TRTTTRYH [1] was used to train SiteGA recognition method [2]. The accuracy (jack-knife) test applied to this dataset have shown that method outperformed conventional dinucleotide position weight matrix (diPWM). Chipmunk approach [2] was applied to deduce the alignment of potential BSs (PBSs) from 4455 sequences (peaks) that contained 15 or more reads in ChIP-Seq annotation [3]. diPWM derived from this alignment was used as the Chipmunk recognition method. The length of matrix 28 nt was selected according to series of jack-knife tests described earlier [2]. The thresholds of recognition methods were computed from EMSA experiment for set of arbitrary selected 64 PBSs. Each PBS (1) was mapped to 1kb long regions upstream transcription starts and (2) had coverage 15 or higher.

Conclusion: Chipmunk and SiteGA methods found PBSs in 78.7 and 76.7% of 4455 peaks, respectively; 88.9% of peaks contained PBSs of at least one method; 56.5% peaks had PBSs of two methods located not farther than 14 nt apart. Only for 8.4% peaks both methods didn't predict PBSs, this portion may represent errors, non-canonical or indirect TF-DNA interactions. Similar analysis of 4367 peaks [5] confirmed these conclusions.

References:

1. V.G. Levitsky et al. (2010) Analysis of data of large-scale chromatin immunoprecipitation by methods of perception of transcription factor binding sites, *VOGiS Herald*, **14**: 685-697.
2. 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.
3. I.V. Kulakovskiy et al. (2010) Deep and wide digging for binding motifs in ChIP-Seq data, *Bioinformatics,* **26**: 2622-2623.
4. E.D. Wederell et al. (2008) Global analysis of in vivo Foxa2-binding sites in mouse adult liver using massively parallel sequencing, *Nucl. Acids Res.,* **36**: 4549–4564.
5. O. Wallerman et al. (2009) Molecular interactions between HNF4a, FOXA2 and GABP identified at regulatory DNA elements through ChIP-sequencing, *Nucl. Acids Res.,* **37**: 7498-7508.

DNA MOTIF SEARCH BY GENETIC ALGORITHM

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

Key words: *DNA motif, genetic algorithm, information content, p-value*

Motivation and Aim: Motif discovery is an important step in annotation of functional sites in the regulatory regions of genes. This task consists in detection of overrepresented conserved sequence patterns in a dataset of functionally related nucleotide sequences. It is supposed that each sequence may contain zero, one or multiple occurrences of binding sites (BS) of any transcription factors (TFs). The position weight/frequency matrix (PWM/PFM) [1] is the most popular method to deduce a motif from an aligned dataset of functional BSs.

Methods and Algorithms: A genetic algorithm (GA) based approach [2] is employed to search for motifs represented as a PWM/PFM of fixed length. It is required that any motif is represented in only a subset of the input dataset (*n* out of total *N* sequences). GA scored motifs based on: (a) KDIC, Kullback-Leibler Discreate Information Content [3]. This value reflects conservation of nucleotides in columns of PFM; (b) p-value, that for given PWM score measures occurrence by chance of sequences that have larger scores. Dependence of p-value from matrix score computed as described in [4].

Results and conclusions: Two type tests were done: (a) alignment of given set of TFBS and (b) motif co-occurrence analysis, i.e. flanking regions of known motifs are searched for other motifs. Developed approach was compared with well-known popular analog MEME [5]. The test (a) revealed that alignment of different tools nearly identical. The test (b) hardly can be adequately interpreted since for analyzed TFBS types actual neighbor motifs are still unknown. Hence, this point requires additional examples, further analysis and biological interpretation. However, our tool and MEME are in good accordance with earlier observation for certain TFBS types (e.g. SF-1, FoxA) that was conserved far beyond known consensuses [2].

References:

1. W.W. Wasserman and A. Sandelin. (2004) Applied bioinformatics for the identification of regulatory elements, *Nat Rev Genet.,* **5**: 276-287.
2. 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.
3. I.V. Kulakovskiy et al. (2010) Deep and wide digging for binding motifs in ChIP-Seq data, *Bioinformatics,* **26**: 2622-2623.
4. H. Touzet and J.S. Varre. (2007) Efficient and accurate P-value computation for Position Weight Matrices, *Algorithms for Molecular Biology*, **2**:15.
5. T.L. Bailey et al. (2009) MEME SUITE: tools for motif discovery and searching, *Nucl. Acids Res.,* **37**: 202-208.

THE ROLES OF THE MONOMER LENGTH AND NUCLEOTIDE CONTEXT OF PLANT TANDEM REPEATS IN NUCLEOSOME POSITIONING

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: levitsky@bionet.nsc.ru
** Corresponding author*

Key words: *nucleosome positioning, tandem repeats, periodicity of dinucleotides*

Motivation and Aim: Tandem DNA arrays are consist of regularly alternating monomers, which have almost identical nucleotide sequences. Such organization makes these arrays especially interesting for clarifying the role of intrinsic DNA preferences in nucleosome positioning (NP). 10-11 –base periodicity of certain dinucleotides is a ubiquitous hallmark of NP.

Methods and Results: We have compared the nucleotide context of the monomers with length (ML) that is equal and unequal to the integer number of the nucleosome repeat length (NRL). 161 plant tandem repeat families from the PlantSat [1] were divided into two classes based on this criterion. We assessed the content of periodic dinucleotides (CPD) in the families of monomers of two classes. The excess of CPD in 1st class was significant according to the chi-square test. Then we applied wavelet transformation [2] to the Phase [3] prediction of NP. The Phase method compares the occurrences of PDs in a potential nucleosomal DNA with those known for approved nucleosomal DNA. Three criteria were used for classification of families of tandem repeats onto types of nucleosome arrangements (NA): (a) the ratio of ML to NRL; (b) the number of peaks in the profiles and their heights; (c) and the heterogeneity of these characteristics within a family. Three main types of NA in DNA tandem repeat arrays were distinguished: Regular (all nucleosomes are positioned in a context-dependent manner), Partial (nucleotide context influences the positioning of only a subset of the nucleosomes), and Flexible (the least effect of nucleotide context in determining NP).

Conclusion: We have demonstrated that the integer ratio of ML to the NRL is accompanied by an increased CPD and a greater influence of the nucleotide context on the NP. Based on the ML and the nucleotide context, three main types of NA in arrays of tandem repeats have been distinguished.

References:

1. J. Macas et al. (2002) PlantSat: a specialized database for plant satellite repeats, *Bioinformatics,* **18**: 28-35.
2. X.Q. Lu et al. (2004) Maximum spectrum of continuous wavelet transform and its application in resolving an overlapped signal, *J Chem Inf Comput Sci*, **44**: 1228-1237.
3. V.G. Levitsky et al. (2008) Nucleosome formation potential estimation via dinucleotide periodicity preferences, *Proceedings of the sixth international conference on bioinformatics of genome regulation and structure,* 140.

INTER-CELLULAR NOISE AND TRANCRIPTIONAL CONTROL OF EPSTEIN-BARR VIRUS LATENCY PROGRAM SWITCHES IN HUMAN B CELL LINES

Li Q., Zou J-Z., Ernberg I.*
Department of Microbiology, Tumor and Cell Biology, Karolinska Institutet, Box 280, Stockholm SE-17177, Sweden
e-mail:Ingemar.ernberg@ki.se
** Corresponding author*

Motivation and aim: Epstein-Barr virus (EBV) is associated with some 200.000 new virus-associated cancer cases/year worldwide. It is thus the 3rd most common tumor virus in man as regards incidence. EBV among human "tumor viruses" is unique in that it is ubiquitous in the human population. More than 90 % of all adults worldwide carry the virus. From this perspective the number of cancer cases are still very low. EBV is involved in the pathogenesis of more than ten different types cancers and lymphomas in man. EBV is the only human tumor virus where the early events in tumor pathogenesis be studied in vitro "under the microscope", hour by hour as the viral transformation of B-cells proceeds. Our studies are focused on the viral control of cellular phenotypic switches, in particular the upstream EBV control of cell proliferation.

Methods: Single-cell data from several thousands of cells in a population can be collected by flow cytometry (FCM) or advanced image analysis (high content screening, HCS). This allowed us to make direct measurements of population variability as well as detection of rare events. We are using two types of closely related human EBV-positive B-cell lines, one expressing the latency I-program (Rael, Mutu I) and the other one expressing latency III-program (CBMI-Ral-Sto, Mutu III).

Results: The virus in latent infection cooperates with cellular transcription and epigenetic factors in controlling switching between cell rest (latency I) and cell proliferation (latency III). It has been known to depend on two viral promoters (Cp,Qp), EBNA 1 and 2. Earlier we have identified cellular regulatory elements of this G0-cell cycle switching: transcription factor OCT2 with its co-regulatory protein Groucho (Grg, TLE). We have now demonstrated that latency 1 to latency III switch can be triggered by downregulation of OCT 2 by shRNA. However, this switch is quite inefficient, and we are therefore exploring additional factors in control of the switch.

Conclusions: The genome-spanning Gene-Regulatory Network (GRN) affords one single genome the capacity to produce a diversity of stable, discretely distinct cell phenotypes in the cellular "state space" that we recognize as "cell types". The distribution of individual cells is mapped in the state-space and we can identify rare cells in one cell type which have acquired properties of the other type. We find evidence for extensive inter-cellular heterogeneity among individual cells in EBV-carrying B cell lines population, presumably reflecting "noise" in the epigenetic programming.

ABOUT SHIFT FUNCTION OF IRREGULAR POLYMERS SYNTHESIS IN MODELS OF THE MATRIX SYNTHESIS

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

Key words: *mathematical modeling, system of ordinary differential equations, variables with delay argument, shift of the function*

Motivation and Aim: The synthesis of DNA, RNA and proteins in living systems is carrying on set stages some of that (elongation of replication, transcription and translation) consist of a lot of intermediate steps linearly following each other. Each stage of elongation is followed only after the initiation stage, which is usually regulated, i.e. is dependent on the concentrations of several substances-regulators. As a result, the elongation stage can be presented by subsystem that obtains some input signal as a control function *f*, processes it and conclude the output signal as a temporal function of target substance's synthesis $p=x_n$. However, some steps of the DNA, RNA (addition of nucleotides to the growing chains) and protein synthesis (joining amino acids to the growing polypeptides) are, in general, not constants, since depend on the concentration of intracellular metabolites and other factors, in turn. In the article we have considered the system:

$$\begin{cases} \dfrac{dx_0(t)}{dt} = \sigma_n(t) f(t) - \dfrac{n}{\tau(t)} x_0(t), \sigma_n(t) = 1 \ \ or \ \dfrac{n}{\tau(t)}, \\ \dfrac{dx_i(t)}{dt} = \dfrac{n}{\tau(t)} \left(x_{i-1}(t) - x_i(t) \right), \ i = 1, \ldots, n-1, \\ \dfrac{dx_n(t)}{dt} = \dfrac{n}{\tau(t)} \left(x_{n-1}(t) - \delta(t) x_n(t) \right). \end{cases} \quad (1)$$

The system is applicable for modeling a wide class of sequential processes, including matrix synthesis of biopolymers such as DNA, RNA and proteins.

Results: ***Limit hypothesis about variable shift of the function.***

The limit hypothesis is formulated for above mentioned system on the next: let $f(t)$, $\alpha(t)$, $\tau(t)$, $\delta(t)$, $\beta_n(t)$ are «enough good» functions (for example, continuous). So, if $\frac{\beta_n(t)}{n} \xrightarrow{n\to\infty} 0$, $\alpha(t)$, $\tau(t)$, $\delta(t)$ – strictly positive functions, the function $\rho(t)$ is a solution of the integral equation $\int_{t-\alpha(t)\rho(t)}^{t} \frac{ds}{\tau(s)} = \alpha(t)$ (2), then for $t > \alpha(t)\rho(t)$ is one of the following limit equalities: (I) if $\delta(t) > 0$, then

$$\lim_{n\to\infty} \frac{n}{\tau(t)\sigma_n(t)} x_{[\alpha(t)n+\beta_n(t)]}(t) = \left[\frac{\tau(t-\alpha(t)\rho(t))}{\tau(t)} \varpi + 1 - \varpi \right] \frac{f(t-\alpha(t)\rho(t))}{\delta(t)}, \varpi = \begin{cases} 1, \sigma_n(t) = 1 \\ 0, \sigma_n(t) = n/\tau \end{cases} \quad (3)$$

There $x_{[\alpha(t)n+\beta_n(t)]}(t)$ is $[\alpha(t)n+\beta_n(t)]$-th component of the Cauchy problem (1) with arbitrary initial data. The article provides a numerical verification of the hypothesis and its proof is given for the particular case.

Conclusion: The hypothesis of limiting the shift function to the variable time delay in the model of the matrix synthesis is formulated. Its numerical verification is conducted and for special cases of limit theorems are proved. The obtained results are new in the theory of biological systems modeling and expand our knowledge about kinetic mechanisms of DNA, RNA and proteins synthesis, functioning of several successive enzymatic reactions, signaling pathways and other dynamic systems, including the extended chains of biochemical processes.

Acknowledgements This study was supported by RFBR (№10-01-00717), Programs of the RAS and SB RAS Presidium (Б.30.29 and 80) and as well as by the scientific school № 5278.2012.4.

GENETIC SUSCEPTIBILITY PROFILE FOR COMORBIDITY VARIANTS OF MULTIFACTORIAL DISEASES

Puzyrev V.P.*[1], Makeeva O.A.[1], Barbarash O.L.[2], Sleptcov A.A.[1], Markova V.V.[1], Polovkova O.G.[1]

[1] *Research Institute of Medical Genetics, SB RAMS, Tomsk, Russia;*

[2] *Research Institute of Complex Issues of Cardiovascular Diseases SB RAMS, Kemerovo, Russia;*

[3] *Genoanalytica, LLC.*

e-mail: valery.puzyrev@medgenetics.ru

** Corresponding author*

Key words: *multifactorial diseases, genetic susceptibility, cardiovascular continuum, syntropy*

Motivation and Aim: The "Cardiovascular Continuum" (CVC) was described in 1991 by Dzau and Braunwald to explain progression of coronary heart disease through other complications and diseases to inevitable end stage of heart failure. Concept of CVC includes several diseases such as coronary artery disease (CAD), arterial hypertension (AH), metabolic syndrome, and diabetes mellitus type 2 (DM2) (Dzau, Braunwald, 1991). In 1921 Pfaundler and von Seht used the term 'syntropy' to designate diseases, which tend to co-occur with each other in patients (or in families) more often than it could be expected by chance (Pfaundler, von Seht, 1921). Based on these two concepts, the term "syntropy genes" (SG) was proposed to designate a set of functionally interacting, co-regulated genes involved in common biochemical and physiological pathways leading to syntopy (Puzyrev, 2008). Thus, the aim was to explore the genetic profile of CVC and to identify SG.

Methods and Algorithms: A large sample of patients with ischemic heart disease and population sample were analyzed to select subgroups for present study. A total of 309 patients out of 800 were selected according to the following criteria: "syntropy" subgroup diagnosed with CAD, AH, DM2, and dyslipidemia in each patient (N=68); comorbidity of CAD and AH with other cardiovascular pathology excluded (n=180), and a subgroup of patients with CAD only (other diseases excluded, N=61). Healthy subjects with normal cardiovascular endophenotypes (N=131) were selected out of a sample of 1600 individuals. Genotyping was done using Illumina Human custom chip microarrays with a panel of markers used for direct-to-consumer genomic service "My Gene" (www.i-gene.ru) (Genoanalytica, LLC). For statistical analysis R v2.14.0 software environment was used, including specialized packages "GenABEL", "snpStats" and "genetics". Predictive value of each candidate SNP was tested using AUC (area under curve).

Results and conclusion: Syntropy group differed significantly from other samples analyzed. Pathway of *ITGA4, KLF7,* and *TAS2R38* genes was involved in the development of this comorbidity. Advanced classifier analysis yielded that *KLF7* rs7568369 reached AUC of 63% and three other SNPs *(LDLR* rs2738446, rs688*,* and *CDKN2A* rs1333048) reached maxAUC of 63%. GG genotype of the rs6501455 located in a region between *KCNJ2* and *SOX9* genes yielded most substantial risk effect in reference to CVC syntropy (OR 3,91; 95% CI 1,56-10,33; $P<0.0016$). GG genotype of the rs7568369 in *KLF7* had highest protective effect in reference to CVC syntropy (OR 0.34; 95% CI 0,16-0,68; $P<7*10^{-4}$). Cluster analysis which involved 90 SNPs, related to different cardiovascular phenotypes, showed that syntopy forms a separate cluster, while other subgroups are close to each other in genetic characteristics.

The study demonstrates that CVC syntropy differs significantly in genetic characteristics from other forms of cardiovascular parthology and has specific genes (SG) involved.

A DRAFT GENOME SEQUENCE OF TARTARY BUCKWHEAT, *FAGOPYRUM TATARICUM*

Logacheva M.D.*[1, 2, 3], Sutormin R.A. [2], Naumenko S.A. [2], Demidenko N.V. [2, 4], Vinogradov D.V. [3], Gelfand M.S. [2, 3], Penin A.A. [2, 3, 4]

[1] *A.N. Belozersky Institute of Physico-Chemical Biology, M.V. Lomonosov Moscow State University, Moscow, Russia;*

[2] *Faculty of Bioengineering and Bioinformatics, M.V. Lomonosov Moscow State University, Moscow, Russia;*

[3] *A.A. Kharkevich Institute for Information Transmission Problems, Russian Academy of Science, Moscow, Russia;*

[4] *Department of Genetics, Biological Faculty, M.V. Lomonosov Moscow State University, Moscow, Russia*

e-mail: maria.log@gmail.com

** Corresponding author*

Key words: *genomics, plant genomes, buckwheat, assembly, annotation*

Motivation and Aim: Despite significant progress in the technologies of DNA sequencing genomic data are lacking for many groups of living organisms, in particular, many plant taxa. Tartary buckwheat (*Fagopyrum tataricum*) belongs to the order Caryophyllales, a large and diverse group of flowering plants, none of which have their genome sequenced and characterized. *F. tataricum* is a close relative of common buckwheat, an important food crop, and a potential donor of favorable traits absent in common buckwheat, such as self-compatibility and improved stress resistance.

Methods and Algorithms: The sequences of five genomic DNA libraries with different insert length (from 100 bp to 3 Kb) and different read length were generated using the Illumina sequencing platform up to 180x genome coverage. In total, about 500 millions of paired-end reads were used for de novo assembly. To facilitate genome annotation and to lay the groundwork for functional genomics we also constructed and sequenced 16 transcriptome libraries corresponding to various stress conditions and developmental stages. They were used for cDNA-based gene prediction and as a training set for ab initio gene prediction.

Results: Draft genome sequence of *Fagopyrum tataricum* was generated. The total number of bases included in the assembly is 372 Mb, that corresponds to 70% of the genome. Nearly 30 thousands of genes were predicted using cDNA-based, homology-based and ab initio approaches. A number of taxon-specific gene duplications, as well as genes specific to the buckwheat genome were found. Using transcriptome sequence data, we also identified alternative transcript isoforms and characterized their expression in different plant organs and under different stress conditions.

Conclusion: We sequenced, assembled and annotated the genome of *F. tataricum* and performed genome-wide analysis of its genes. Given the close relationships of *F. tataricum* to cultivated buckwheat we expect that this sequence will be useful not only for evolutionary and comparative genomics, but also for practical aspects such as identification of genes responsible for agriculturally important traits.

COMPUTING DNA OLIGONUCLEOTIDES HYBRIDIZATION ENTHALPY WITHIN MOLECULAR DYNAMICS MODELING

Lomzov A.A.*, Vorobjev Y.N., Pyshnyi D.V.
Institute of Chemical Biology and Fundamental Medicine, SB RAS, Novosibirsk, Russia
e-mail: lomzov@niboch.nsc.ru
** Corresponding author*

Key words: *enthalpy, DNA, hybridization, Molecular dynamics, GPU*

Motivation and Aim: Development of new derivatives and analogues of nucleic acids (NA) and reliable prediction of their physico-chemical properties is important both in practice and basic research. Significant progress in development of software and hardware has made the *in silico* research widely used. The goal of this work is to analyze an applicability of the molecular dynamics (MD) modeling for calculating oligonucleotide hybridization enthalpy.

Methods and Algorithms: The enthalpies of DNA duplex formation were determened as a difference of the internal energy of double- and single-stranded states which were calculated from MD trajectory computed with Amber 11 software (UCSF, USA). Computations were performed on NVIDIA GTX580 and Intel i7-2600 hardware.

Results: To determine optimal parameters of modeling we have used Dickerson-Drew dodecamer (DDD) with well characterized secondary structure and thermal stability. We have varied temperature, heating protocol, and ion concentration in implicit and explicit solvent and compared averaged structures with those experimentally obtained. Using optimal parameters of modeling we have shown that hybridization enthalpy of DDD correlates well with experimental and calculated one via nearest neighbor model enthalpies. The use of GPU has speeded up the modeling of DDD in implicit solvent up to 60 times and up to 30 in explicit solvent in comparison with the single node CPU.

To verify the MD predictive ability we have collected database of experimentally determined thermodynamic parameters (enthalpy, entropy and Gibbs energy) of hybridization of 272 oligodeoxyribonucleotides. The total energy of oligonucleotide and its complex were calculated from 2 ns trajectories simulated with optimal parameters. The RMSD of calculated and experimental enthalpies was 15%.

Conclusion: The results obtained show that MD modeling allows one to calculate enthalpy of matched DNA duplexes with high accuracy.

Availability: An extension of this work is retrieval of parameters of MD modeling for more accurate prediction DNA duplex thermal stability, including complexes with perturbation of the regular structure, that could be used instead experimental research.

This research has been supported by Integration grant SB RAS (86), RFBR (10-04-01492-a) and by MCB programs of RAS.

MODELING RNA POLYMERASE INTERACTION IN PLASTIDS OF PLANTS, ALGAE AND MITOCHONDRIA OF CHORDATES: HUMAN BEARING THE MELAS MUTATION AND RAT WITH HYPOSECRETION OF THYROID HORMONE

Lyubetsky V.A.*, Seliverstov A.V.
Institute for Information Transmission Problems of the Russian Academy of Sciences (Kharkevich Institute), Moscow, Russia
e-mail: lyubetsk@iitp.ru
** Corresponding author*

Key words: *RNA polymerase interaction, plastids, mitochondria, plants, chordata*

Motivation and Methods: We introduced a concept, a mathematic model and its computer realization that describe the interaction between bacterial and phage type RNA polymerases, protein factors and secondary structures during transcription, including transcription initiation and termination. The model accurately reproduces virtually all relevant experimental data available on plastids of plants and algae, and mitochondria of chordates (frog, rat and human). The model was shown to accurately reproduce changes of gene transcription level observed in polymerase sigma-subunit knockout and heat shock experiments on plastids of plants and algae; and most evidence on bulk RNA contents and RNA half-life times in mitochondria of frog, healthy human, human bearing the MELAS mutation, healthy rat, and rat with hyposecretion of thyroid hormone. Predicted transcription characteristics are: percentage of polymerases terminated in both directions at a protein-dependent terminator; binding intensities of the regulatory protein factor (mTERF) with the termination site; transcription initiation intensities of all promoters in three chordate species (frog, healthy human, human with MELAS syndrome, healthy rat, and hypothyroid rat with aberrated mtDNA methylation). Absolute levels of gene transcription are obtained, while only relative RNA contents are known for selected genes from the experiment.

Results: A model was introduced to describe the interaction between moving ribosomes on RNA (a polysome) and ribonucleases. We identified putative factors mediating the MELAS syndrome development in human: the decrease of Phe-tRNA, Val-tRNA and rRNA contents in the cell. In human with MELAS syndrome the model predicts the noticeable 1.21-fold decrease of the mTERF-DNA binding intensity and the 7.75-fold decrease of the HSP1 promoter efficiency. Transcription levels of tRNA-Phe and rRNA drop 3.84- and 1.2-fold, respectively, that suggests possible implications for the MELAS phenotype. Intensities of the mTERF binding and LSP transcription initiation are equal between eu- and hypothyroid rats. The total intensity of transcription initiation from promoters HSP1 and HSP2 is 2.15-fold lower in the hypothyroid, which conforms well to the experimentally known methylation patterns of relevant DNA loci in eu- and hypothyroids. We describe the correlation between changes in methylation patterns of the mTERF binding site and three promoters in hypothyroid rat, and between changes in intensities of the mTERF binding and transcription initiations.

Availability: The corresponding computer program and a user guide are available at http://lab6.iitp.ru/en/rivals.

IN SILICO STRUCTURAL 3D MODELLING OF NOVEL *CRY1I* AND *CRY3A* GENES FROM LOCAL ISOLATES OF *BACILLUS THURINGIENSIS*

Mahadeva Swamy H.M. [1], Asokan R. [1], Mahmood R.[2]
[1] *Division of Biotechnology, Indian Institute of Horticultural Research (IIHR), Hessarghatta lake post, Bangalore 560089 INDIA;*
[2] *Post-Graduate Department of Studies and Research in Biotechnology and Bioinformatics, Kuvempu University, Jnanasahayadri, Shankaraghatta, Shimoga 577451, Karnataka, INDIA*
Corresponding author e-mail: clintonbio@gmail.com

Key words: *3D models, homology modeling, coleopteran insects, domains*

Motivation and aim: Determining the structure and function of a novel protein is a cornerstone of many aspects of modern biology.

Methods and algorithms: The 3D structures was predicted using phyre2 server (http://www.sbg.bio.ic.ac.uk/phyre2/), Conserved Domains and Protein Classification (http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml, version CDDv2.32-40526 PSSMs) and the predicted structure was validated using protein structure validation software suite (PSVS) tool. Determination of protein functional analysis obtained from databases using InterProScan (http://www.ebi.ac.uk/Tools/pfa/iprscan/) and ProFunc (http://www.ebi.ac.uk/thornton-srv/databases/profunc/index.html).

Results: Three-dimensional (3D) models for the 79.2-kDa activated *Cry*1I and 77.4kDa activated *Cry*3A δ-endotoxins from *Bacillus thuringiensis* native isolates that are specifically toxic to Coleopteran insect pests were constructed by homology modeling. They were structurally similar to the known structures, both derived 3D models displayed a three domain organization: the N-terminal domain (I) is a seven helix bundle, while the middle and C-terminal domains are primarily comprise of anti-parallel β-sheets. Significant structural differences within domain II in this model among all Cry protein structures indicates that it is involved in recognition and binding to cell surfaces. Comparison of Coleopteran-active cry toxins predicted structure with available experimentally determined Cry structures reveals identical folds.

Conclusions: The collective knowledge of Cry toxin structures will lead to a more critical understanding of the structural basis for receptor binding and pore formation, as well as allowing the scope of diversity to be better appreciated. Taken together, these studies provided promising evidence that domain swapping, epitope-mapping and protein-engineering under the guidance of molecular modeling can serve as a rational and useful tool in understanding the mode of action of Cry toxins, and ultimately in producing better toxins. Structural insights from these molecular modeling studies would therefore increase our understanding of the mechanic aspects of these two closely related Coleopteran-active insecticidal crystal proteins. These proteins are of interest for agriculture, as they offer a means for control of beetles and other insect crop pests.

Availability: Academic, Acknowledgments: The authors are grateful to Indian Council of Agricultural Research (ICAR), New Delhi for funding this study under Network project on Application of Microbes in Agriculture and Allied Sectors (AMAAS). Infrastructure facility and encouragement by The Director, Indian Institute of Horticultural Research (IIHR) are duly acknowledged.

EVOLUTION OF NON-CODING MITOCHONDRIAL SEQUENCES OF THE BAIKALIAN SPONGES (LUBOMIRSKIIDAE)

Maikova O.O.*, Sherbakov D.Y., Belikov S.I.
Limnological Institute, SB RAS, Irkutsk, Russia

e-mail: maikova@lin.irk.ru
** Corresponding author*

Motivation and Aim: The nucleotide sequence of mtDNA has been successfully used as a molecular marker for resolve of phylogenetic and evolutionary tasks. Mitochondrial DNA evolution of sponges, the basal group of animals on the phylogenetic tree of metazoans, arouses particular interest to scientists. Recently, the low rate of evolution of mtDNA coding sequences of sponges relatively more advanced animal groups has shown. In other animal groups (cnidarians, insects) non-coding mtDNA sequences are used to determine interspecies and population boundaries. Previously in non-coding mtDNA sequences of Baikal sponges many inverted repeats were found and their mutation rate four times the rate of single nucleotide substitutions (Lavrov, 2010). In the present study we examined the evolution of the intergenic spacers of Baikal endemic sponge mtDNA (family Lubomirskiidae).

Methods and Algorithms: Search and comparative analysis of secondary structures (inverted repeats) was performed using the EMBOSS palindrome. The pseudoknot sequences were identified using the pknotsRG 1.3 program according to loc-strategy. Statistical analysis of possible recombination events between the complete mtDNA genome sequences of four species of Baikal sponges was performed using the PhiPack and TOPALi programs. Phylogenetic Bayesian inference was conducted based on the mtDNA non-coding sequences, which were aligned according to their secondary structure using locARNA (Freiburg RNA Tools).

Results: Comparative and phylogenetic analyses based on mtDNA non-coding sequences of Baikal sponges have shown, that pattern of inverted repeats are specific for the species but sometimes we observed changes of the pattern within species. On the phylogenetic tree the *B. recta* and *B. martinsoni* species form a mixed cluster. The sponges of *B. intermedia profundalis*, *B. intermedia*, *B. bacillifera* and *L. baicalensis* species were grouped into the separate clusters, but the two *L. baicalensis* were grouped separately from other specimens of this species, together with *B. bacillifera*.

Conclusion: Based on these results, we suggested a partial separation of ancestral lineages of the mitochondrial haplotypes of non-coding sequences of Baikal sponges, therefore the morphologically different species inherited genetic polymorphism from their ancestors. Possibly, however, that these phenomenon are due to the genetic transfer between Baikalian sponges species such as horizontal transfer of DNA, for example. In favor of the second assumption there is the fact that with high probability (Phi=0,0693) the asymmetric horizontal transfer of non-coding fragments of mtDNA was possible between species of Baikal sponges. Presence of pseudocnot sequences in Baikal sponge mtDNA which were found in the intergenic regions also confirms our hypothesis, their existence indicates a high probability of intramolecular recombination. As well as a possible mechanism for the proliferation of inverted sequences could be polimerase slipped-strand mispairing.

Availability: All programs are available over the Internet: EMBOSS (emboss.sourceforge.net), pknotsRG 1.3 (http://bibiserv.techfak.uni-bielefeld.de/pknotsrg/), TOPALi (http://www.topali.org/), locARNA (http://rna.informatik.uni-freiburg.de:8080/LocARNA.jsp).

References:

1. 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.

IDENTIFICATION OF THE KEY COMPONENTS TO CONTROL THE BEHAVIOUR OF A COMPLEX PATHWAY: A STUDY ON A MODEL OF MITOCHONDRIAL BIOENERGETICS

Maj C.[1,2], Mosca E.[1], Merelli I.[1], Mauri G.[2], Milanesi L.*[1]
[1] *Institute for Biomedical Technologies CNR, Via Fratelli Cervi, Segrate, Milano, Italy;*
[2] *University of Milano-Bicocca, DiSCo V.le Sarca 336, Milano, Italy*
e-mail: luciano.milanesi@itb.cnr.it
** Corresponding author*

Key words: *systems biology, mathematical model, sensitivity analysis*

Motivation and Aim: An important aim in systems biology is to understand how to predict and control the complex dynamics of biological pathways. In this work, we tackle this issue considering a particular type of sensitivity analysis, called regionalised sensitivity analysis (RSA) [1]. We use this approach to study transporters and enzymes involved in the regulation of [ATP]/[ADP] ratio (r) in a detailed kinetic model (differential algebraic equations) of mitochondrial bioenergetics [2].

Methods and Algorithms: We generated, using quasi-random sequences, 10,000 realizations of the 42-dimensional vector of the model kinetic parameters, where each parameter, which represent the activity of a protein (transporter or enzyme), was increased or decreased within two orders of magnitude above and below its original value. We simulated the model using these 10,000 configurations and each kinetic parameters realization was classified into one of two different sets, considering whether r was or not greater than 1. Then, we estimated the two distributions of each parameter values in both these groups and defined as influential the parameters having significantly different distributions between the two sets (two sample Kolmogorov-Smirnov test, corrected for multiple testing).

Results: Most of the model configurations screened (85.9%) lead the model into states where $r > 1$. A total of 14 out of the 42 enzymes or transporters were found to have a significant control ($p < 0.05$) over r. At the top of the list we found the adenine nucleotide translocase ($\log_{10}(p) = -138$), ion transporters and respiratory complexes. The number of simulations carried out was sufficient to ensure the convergence of the final list of proteins that significantly control r.

Conclusion: RSA is a suitable and effective approach to study the dynamics of a complex model representing a biological pathway. RSA is based on a high number of model simulations and thus it requires high performance computing.

Acknowledgements: HPC-Europa2 project (228398), Italian FIRB-MIUR project HIRMA, Flagship Initiative INTEROMICS; CM is a fellow of the PhD programme of the University of Milano-Bicocca.

References:

1. Saltelli A, Ratto M, and Andres T. Wiley Online Library, 2008.
2. Bazil JN, Buzzard GT, Rundell AE (2010). PLoS Comput Biol 6(1).

UNSUPERVISED ALGORITHM BASED ON WAVELET ANALYSIS FOR EXTRACTION OF INFORMATION ABOUT HIPPOCAMPAL NEURONAL ACTIVITY CHARACTERISTICS FROM THE EXPERIMENTAL DATA

Malakhin I.A.
The Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia;
Design Technological Institute Of Digital Techniques, SB RAS, Novosibirsk, Russia
e-mail: Pilat1988@gmail.com

Key words: *field action potential, unsupervised analysis, wavelet analysis*

Motivation and aim: Long-term and costly experiments require the fullest extraction of useful information from the experimental data. This process remains protracted even if the well-known semiautomatic programs were used.

The purpose of the work is to develop an algorithm and its program realization that: firstly, provides unsupervised information extraction and secondly increases the number of extracted characteristics.

Methods and Algorithms: Multilevel discrete wavelet decomposition is the key method used in the algorithm [1]. The records of evoked field action potentials (fAP) of mice hippocampus pyramidal neurons obtained by different stimulation protocols in different conditions were analyzed.

Results: The algorithm of unsupervised searching of meaningful patterns in experimental data has been developed. The effective noise filtration and finding the frequency pattern related to biological events has been realized by using the wavelet decomposition. There are two key differences between the program realization of this algorithm and the most existing programs for electrophysiology data analysis. The first difference of the developed program is a lack of the step when a researcher has to configure the program and set the region of interest. The second, finding the fAP is based on finding the extrema rather than on threshold intersection. The program realization of the algorithm makes it possible to increase the number of extracted characteristics from 2 to 13 and decrease the analysis time two orders of magnitude as compared with the semiautomatic method using before.

Conclusion: The program realization of the developed algorithm makes it possible to increase the precision of extraction and the number of characteristics extracted from experimental data.

Availability: the program source code is published under GPLv3 license and is available at https://c-fos@github.com/c-fos/fEPSP-analyser.git.

Acknowledgements: The work was supported by RAS base fundamental research project VI.53.1.3, Integration project presidium SB RAS № 136, RFBR grant № 12-01-00639.

References:

1. A.B.Wiltschko at al. (2008) Wavelet filtering before spike detection preserves waveform shape and enhances single-unit discrimination. *Journal of Neuroscience Methods* **173**: 34-40.

THEORETICALLY-EXPERIMENTAL RESEARCH OF VESICLE TRAFFICKING MECHANISMS IN THE SYNAPTIC PLASTICITY PROCESS

Malakhin I.A.[1, 2], Proskura A.L.[1], Vechkapova S.O.[2], Zapara T.A.*[1], Ratushniak A.S.[1]
[1] *Design Technological Institute of Digital Techniques, SB RAS, Novosibirsk, Russia;*
[2] *The Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia*
e-mail: zap@kti.nsc.ru
** Corresponding author*

Key words: *LTP, vesicle trafficking, glutamate receptors, synaptic plasticity*

Motivation and Aim: In the process of expression and maintenance the long-term changes of structure-functional state of dendritic spine more than 2000 various proteins and their interactions regulate unknown mechanisms of equilibrium between plasticity and stability in the molecular system. On the base of integration and analysis of experimental data the molecular processes responsible for changing and maintenance the number of AMPAR [AMPA (alpha-amino-3-hydroxy-5-methylisoxazole-4-propionic acid)] on synaptic membrane has been reconstructed. The protein-protein network "Vesicle trafficking of mediator receptors" describing the main steps of delivering AMPAR in synapse has been developed. The analyses of this theoretical model permitted to make a conclusion that the integrity of vesicle pathway is all-important for the process of neuronal plasticity. On the base of these data the task has been formulated and the experimental verification of this hypothesis has been made.

Methods and algorithms: To reconstruct the model of molecular network the GeneNet technology was used. The experiments were performed on the hippocampal slices of mice of ICR line using the model of long-term posttetanic potentiation and standard electrophysiological methods.

Results: The theoretical analysis of mechanisms that control the delivery of AMPAR into synapse has shown that long-term potentiation (LTP) expression depends on quick incorporation of GluR1-AMPAR into synaptic membrane and recycling of GluR2-AMPAR. The LTP maintenance depends on the delivering a new synthesized AMPAR into a synapse. The analysis of the network permitted to develop a scheme of experiment for investigating the influence of Brefeldin A, a vesicle trafficking blocker, on LTP expression and maintenance. A series of electro-physiological experiments was set on the mice hippocampal slices. It has been shown that Brefeldin A in the concentration 1 mkg/ml and 5 mkg/ml prevents LTP maintenance and has not effected on its expression. In lower concentrations (0.1 mkg/ml and 0.03 mkg/ml) Brefeldin A did not influence on LTP expression or maintenance.

Availability: available on request from the authors.

Conclusion: On the base of analysis of reconstructed network we managed to formulate and experimentally verify the hypothesis that delivery of vesicle with a new synthesized AMPAR maintain the synaptic conductivity and synapse efficiency.

Acknowledgements: The work was supported by RAS base fundamental research project VI.53.1.3, Integration project presidium SB RAS № 136, RFBR grant № 12-01-00639.

TOWARDS AN UNDERSTANDING OF THE ROLE OF HUMAN RIBOSOMAL PROTEINS IN VARIOUS CELLULAR PROCESSES RELATED TO HEALTH AND DISEASES

Malygin A.A.*, Ivanov A.V., Babaylova E.S., Karpova G.G.
Institute of Chemical biology and Fundamental Medicine SB RAS, Novosibirsk, Russia
e-mail: malygin@niboch.nsc.ru
** Corresponding author*

Key words: *human ribosomal proteins, gene expression regulation, HCV, IRES*

Motivation and Aim: Ribosomal proteins are constitutive components of the ribosome and, therefore, they are involved in the work of translation machinery. However, apart from their "ribosomal" functions, many ribosomal proteins are also implicated in a variety of other cellular processes related to health and disease.

Results: We found that human ribosomal protein (rp) S13 can regulate expression of its own gene at the splicing step by a feedback mechanism. According to this mechanism, rpS13 was demonstrated to inhibit excision of intron 1 from its own pre-mRNA due to the binding to the intron nearby the splicing sites that interferes interaction of the conventional splicing factors with these sites [1]. The resulting mRNA retains the intron and therefore is nonsense. Similar mechanism was found to be realized with rpS16 [2] and rpS26 [3]. The feedback regulation of ribosomal protein genes expression at the splicing step may provide fine tuning of the level of each ribosomal protein in cell. Ribosomal protein SA is known not only as a component of the ribosome, but also as a precursor of the cell-surface laminin receptor, LAMR. RpSA is homologous to eubacterial rpS2, but in contrast to it rpSA is not a constant ribosomal component. It has a eukaryote-specific C-terminal domain that was found to be responsible for the protein binding to the 40S ribosomal subunit involving mainly the 18S rRNA helix 40. The C-terminal domain of rpSA was shown to contain a receptor domain for Venezuelan equine encephalitis and tick-borne encephalitis viruses. We examined the internal ribosome entry site (IRES) of the hepatitis C virus (HCV) RNA for its ability to bind to 40S ribosomal subunits (i) deficient in rpSA, (ii) saturated with recombinant rpSA, or (iii) pretreated with monoclonal antibodies whose epitops are located in the C-terminal domain of rpSA. Binding of HCV IRES to 40S subunits was shown to depend largely on the rpSA content in the subunits and to be blocked by the antibodies [4]. The results obtained imply that eukaryote-specific C-terminal domain of rpSA is implicated in binding of the HCV IRES to the ribosome and, therefore, in translation initiation of HCV RNA.

Acknowledgements: The work was supported by grant from RFBR 11-04-00672-a.

References:

1. A.A. Malygin et al. (2007) Human ribosomal protein S13 regulates expression of its own gene at the splicing step by a feedback mechanism, *Nucleic Acids Res.,* **35:** 6414-6423.
2. A.V. Ivanov et al.. (2010) Human ribosomal protein S16 inhibits excision of the first intron from its own pre-mRNA, *Molecular Biology,* **44:** 82-88.
3. A.V. Ivanov et al. (2005) Human ribosomal protein S26 suppresses the splicing of its pre-mRNA, *Biochim. Biophys. Acta,* **1727:** 134-140.
4. A.A. Malygin et al. (2009) Binding of the IRES of hepatitis C virus RNA to the 40S ribosomal subunit: role of p40, *Molecular Biology,* **43:** 997-1003.

SPATIALLY DISTRIBUTED MODELING OF BACTERIAL COMMUNITIES WITH HAPLOID EVOLUTIONARY CONSTRUCTOR

Mamontova E.A., 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*

Motivation and Aim: Complex multiparameter computer models of evolution become more and more powerful tool for the analysis of evolutionary, population genetics, and ecological hypotheses. Since its inception, models of evolution complicated both in terms of methods used and problems to be solved. The "Evolutionary constructor" is a recently developed methodology for constructing and analyzing the multi-layer models of evolutionary processes [1, 2]. Its software implementation "Haploid evolutionary constructor" (HEC) is specified to simulate bacterial communities living in bounded flowing environment. Community members interact via trophic relations: substrates, which are synthesized and secreted by one population of cells, can be consumed by cells of another population. The processes of substrates metabolism in cells are regulated by related genes on the principle "one gene – one constant of utilization/synthesis". The current HEC version does not support spatial distribution of substrates and cells in an environment, which reduces the biological sense of simulation results. Thus our goal was to improve HEC by developing the additional program modules providing spatial distribution in 1D, 2D and 3D.

Methods and Algorithms: In order to model spatial distribution we divided the whole environment on subareas (1D, 2D, 3D, respectively) and described the flows and diffusion between them using standard equations. The migration of cells between subareas was described using genetic spectra arithmetics [1]. The program modules were written in C++, as the HEC computational core is written in it.

Results and conclusion: The methods for describing the spatial distribution of substrates and cells in the HEC models were developed. The software implementation is complete for 1D case (2D and 3D are being developed). It opens users the new opportunities to simulate bacterial communities closer to reality.

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

Acknowledgements: The work was supported by the RFBR grants 10-04-01310-a, 12-07-00671-a, RAS Program № 28.

References:

1. S.A. Lashin et al. (2010) Comparative modeling of coevolution in communities of unicellular organisms: adaptability and biodiversity, *JBCB*, **8:** 627-643.
2. S.A. Lashin et al. (2012) Computer modeling of genome complexity variation trends in prokaryotic communities under varying habitat conditions, *Ecol. modelling,* **224:** 124-129.

VALIDATION OF AFFYMETRIX PROBE SETS: NEW APPROACHES TO THE OLD PROBLEM

Marakhonov A.V. * [1], Sadovskaya N.S.[1], Baranova A.V.[1, 2], Skoblov M.Yu. [1, 3]
[1] *Federal State Budgetary Institution "Research Centre for Medical Genetics" under the Russian Academy of Medical Sciences, Moscow, Russia;*
[2] *School of Systems Biology, College of Science, George Mason University, Fairfax, VA USA;*
[3] *State Budgetary Institution of Higher Education "Moscow State Medical and Dental University", Moscow, Russia*
e-mail: marakhonov@generesearch.ru
** Corresponding author*

Key words: *gene expression, microarray, validation, probe set*

Motivation and Aim: Recent advances in methodology have led to the extent studies of organisms at the level of nucleic acids. Microarray gene expression profiling offers an opportunity for genome-scale, quantitative evaluation of gene expression studies by simultaneously measuring expression levels for thousands of genes. Nevertheless the discrepancies in probe sets data remained to be wide-spread phenomenon which is usually underestimated or dissembled. So the correct annotation of probe sets remains the actual problem. The man goal of the study was to determine probe sets which could correctly represent the real gene expression profile in human transcriptome.

Methods and Algorithms: We choose averaged large microarray data set GSE1133 from BioGPS [1, 2]. The data sets were originally generated using Affymetrix arrays (U133 Plus 2.0 and U133A platforms). We extracted the genes containing more than one probe sets. After that we performed correlation analysis of probe sets to each gene. We have compared probe sets to each gene with PLANdbAffy database focused on the probe-level annotation for Affymetrix expression microarrays at a level of sequence [3].

Results: The performed analysis allowed us to distinguish the multiple probe sets containing genes into several groups. Highly correlated probe sets to the single genes could indicate the specific probe sets which are correctly annotated. Two other types of uncorrelated probe sets could be related to the different transcript variants and tissue-specific genes. The one of cause of uncorrelated data from probe sets to the single genes remained to be improper annotation of probe sets.

Conclusion: Wide-spread existence of uncorrelated probe sets to the single genes indicated the mismapping or the misannotation of these probe sets. Furthermore the used approach allowed determining group of tissue-specific genes and tissue-specific probe sets which could correspond to the novel tissue-specific transcript variants of the genes.

References:

1. C. Wu et al. (2009) BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources, *Genome Biol,* **10:** R130.
2. A.I. Su et al. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes, *Proc Natl Acad Sci U S A,* **101:** 6062–6067.
3. R.N. Nurtdinov et al. (2010) PLANdbAffy: probe-level annotation database for Affymetrix expression microarrays, *Nucleic Acids Res,* **38:** D726–30.

ADVANCES IN GENOMIC AND METAGENOMIC STUDIES OF EXTREMOPHILIC MICROORGANISMS

Mardanov A.V., Kadnikov V.V., Gumerov V.M., Ravin N.V. *
Centre "Bioengineering" RAS, Moscow, Russia
e-mail: nravin@biengi.ac.ru
** Corresponding author*

Key words: *extremophiles, archaea, genome, microbial community, metagenomics*

Sequencing of microbial genomes is one of the departure points for researches in the fields of microbiology, molecular biology and evolution of living organisms, and also has practical importance for medicine and biotechnology. The objects of our research are microorganisms living in extreme environments, primary thermophilic archaea. They are evolutionary ancient organisms with relatively small genomes but possessing complete systems of autotrophic or heterotrophic metabolism. We determined complete genome sequences of twelve thermophilic microorganisms representing either new phylogenetic lineages or biotechnologically/environmentally important species. Genomic data were used for identification of the metabolic pathways of investigated microorganisms, studying the molecular mechanisms of genetic processes such as DNA replication, analysis of evolution of the genome and individual genes, structural and functional studies of particular proteins. Next generation sequencing technologies facilitated analysis of the structures of microbial communities based on deep sequencing of 16S RNA clones and studies targeted to the whole community metagenome. We investigated metagenomes of extreme environments, - methane hydrate bearing sediments of the Lake Baikal, deep subsurface thermal waters in Tomsk region and hot springs of Kamchatka region. The results obtained revealed the influence of extreme conditions on the diversity of microorganisms, the main biogeochemical processes in these extreme environments, and new groups of prokaryotes representing deep phylogenetic lineages.

DEVELOPMENT OF THE OPTIMAL ALGORITHM OF BACTERIAL WHOLE GENOME SEQUENCING ON MISEQ AND GS JUNIOR 454 SEQUENCERS

Markelov M.L., Gordukova M.A.*, Kuleshov K.V., Dedkov V.G., Alvarez Figueroa M.V.
FBIS Central Research Institute for Epidemiology, Moscow, Russia
e-mail: maria.gordukova@pcr.ru
** Corresponding author*

Motivation and Aim. We deal with two types of tasks for whole genome analysis: resequencing with analysis of SNPs and *de novo* assemble. Precision of SNPs analysis is important for interlaboratory networks for the investigation of outbreaks and clinical tasks related to the detection of mutations linked to microorganisms' resistance. We have evaluated relative performance of new platforms with the aim to offer the optimal algorithm for bacterial whole genome sequencing. We used two Personal Sequencing Platforms: GS Junior System 454 (Roche Diagnostics) and MiSeq (Illumina). We performed whole genome sequencing of two different clinically significant bacterial strains with different GC-content: the vaccine strain of *M. bovis* BCG Russia (65.4%) (BCG) and *V. cholerae* (47.6%) (VCh).

Methods and Algorithms. Both genomes were sequenced using two runs on 454 to get enough coverage for high quality *de novo* assemble. For MiSeq sequencer we realized one run with pooled libraries from BCG and VCh. For *de novo* assemble of 454 data we used Newbler v2.5 software, for MiSeq pair-end reads – CLC Genomics Workbench v5.0.1. To make investigation and correction of SNVs and indels appearing due to low coverage of *de novo* contigs obtained from 454 we mapped MiSeq pair-end reads to 454 *de novo* assembled contigs.

Results. We received 112 Mb for BCG and 87 Mb for VCh from 2 runs of each sample by using 454 and 550Mb for BCG and 508.2 Mb for VCh with pooled libraries by using MiSeq. Some data on sequencing results and *de novo* assembly contigs for each bacterial strain are summarized in the table.

Strain	Sequencer	Runs	Sequenced reads	Average length of reads	Average coverage of *de novo* contigs	*De novo* contigs	N50, kb
BCG	MiSeq	0.5	3 741 498	147x2	98	243	38
	454	2	275 865	477	27	146	101
Vch	MiSeq	0.5	3 741 498	147x2	102	202	55
	454	2	218 882	350	33	124	135

We detected incorrectly identified nucleotides and indels in homopolymer regions in VCh sequenced contigs and single nucleotide deletions in GC-rich regions from contigs of BCG strain by mapping MiSeq reads on contigs generated with 454.

Conclusion. On the basis of these data we offer the following algorithm of genome sequencing of bacteria and *de novo* genome assembly. If necessary to receive long contigs, the first step would be to use 454 due to its ability to generate long reads to obtain long contigs for subsequent scaffolding procedures. Further due to the previously described limitations of pyrosequencing technology when dealing with homopolymers it is necessary to confirm the accuracy of the received sequences. It is only possible with huge coverage reads through Sequencing-by-Synthesis usingIllumina instrument.

CORRELATION BETWEEN TRANSCRIPTION EFFICIENCY INITIATION AND TRANSLATION EFFICIENCY FOR *SACCHAROMYCES CEREVISIAE* AND *SCHIZOSACCHAROMYCES POMBE*

Matushkin Yu.G.*, Levitsky V.G., Orlov Y.L., Likhoshvai V.A.
Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia;
Novosibirsk State University, Novosibirsk, Russia
e-mail: mat@bionet.nsc.ru
** Corresponding author*

Key words: *translation efficiency, nucleosome formation potential, elongation efficiency index correlations, Saccharomyces cerevisiae, Schizosaccharomyces pombe*

Motivation and Aim: The hypothesis to test is the following: effective gene expression needs coordinately optimized processes of transcription and translation, in particular – transcription initiation and translation elongation.

Methods and Algorithms: Elongation Efficiency Index (EEI) was suggested earlier in [1] to estimate gene expression efficiency by nucleotide context of coding sequence in unicellular organisms. We have analyzed association between EEI and nucleosome formation potential (NFP) in 5' regulatory regions upstream translation initiation site (TIS) from two yeast species. Theoretical estimations of NFP based on DNA sequence were obtained by Recon method [2].

Results: Elongation efficiency negatively correlates with nucleosome formation potential. Therewith the selection may lead both on NFP decrease for high-expressing sequences (for *S.pombe*), and on NFP increase for low-expressing ones (for *S. cerevisiae*). Apparently this is the cause of distinct distributions of poly(A) tracts in upstream regions for two yeast species. For both yeast species there are regions with significant negative correlation of elongation efficiency index and nucleosome formation potential in 5' regulatory regions for all genes and especially for high expression level genes estimated by EEI. This negative correlation is more significant for *S.pombe.* Therefore for genes with low expression level there are regions with significant positive correlation, and this positive correlation is more significant for *S.cerevisiae*. In addition, for *S.cerevisiae* there is strong significant correlation between elongation efficiency index and nucleosome formation potential in coding regions for all genes and especially for genes with high expression level.

Conclusion: We have shown inter-relation between nucleosome localization in promoter regions and elongation efficiency index for yeast species *S. cerevisiae* and *S. pombe*.

References:

1. N.V. Vladimirov, V.A. Likhoshvai, Yu.G. Matushkin (2007) Correlation of Codon Biases and Potential Secondary Structures with mRNA Translation Efficiency in Unicellular Organisms. *Mol Biol (Mosk)*., **41**(5): 843–850.
2. Victor G. Levitsky (2004) RECON: a program for prediction of nucleosome formation potential. *Nucleic Acids Res*., **32: 346–349.**

STATISTICAL ANALYSIS OF DATABASE DERIVED INTER-RESIDUE CONTACT POTENTIALS

Mavropulo-Stolyarenko G.R.
Saint-Petersburg State University, Saint-Petersburg, Russia
e-mail: gm2124@mail.ru

Key words: *protein structure prediction, comparative modelling, statistical contact potentials*

Motivation and Aim: So-called, statistical contact potentials, have proven to be a valuable tool of modern days protein structure modelling. By applying Boltzmann formalism, to probabilities of amino acids being in contact in real protein structures and in random "reference state", energy-like measures are derived that allow filtering decoy from a native-like structures [1,2]. Thus our understanding of how to construct more precise and sensitive potentials is the key to a possible breakthrough in protein structure prediction. The purpose of this study is to statistically assess "database derived contact potentials" model in order to find ways to improve it.

Methods and Algorithms: We had performed a statistical analysis of distance dependent contact potentials derived from the fresh (PDBselect 25% Mar. 2012) non redundant protein structure list [3]. As a random reference state, both: contact map (amino-acid sequence) shuffle and Monte Carlo random structure generation models were implemented. We used binomial distribution statistics with multiple comparisons correction for testing contact probabilities for equality.

Results: Careful statistical analysis of statistical contact potentials had shown that most of the currently known methods to construct them, oversimplify underlying model thus missing some of its statistically significant parts. Although it is hard to estimate the amount of the error introduced by these simplifications, our findings suggest that existing statistical potentials could be improved.

Conclusion: Results presented in this study show possible paths of improving statistical potentials, to account for newly discovered effects like N-C pair asymmetry or low intra-sequence separation.

Availability: software used in this research will be available on request from the authors since the fall of 2012;

References:

1. Hendlich M, Lackner P, Weitckus S, Floeckner H, Froschauer R, Gottsbacher K, Casari G, Sippl MJ. Identification of native protein folds amongst a large number of incorrect models. The calculation of low energy conformations from potentials of mean force. J Mol Biol. 1990; 216:167–80.
2. Sippl MJ. Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures. J Comput Aided Mol Des. 1993; 7:473–501.
3. Sven Griep and Uwe Hobohm. PDBselect 1992–2009 and PDBfilter-select. Nucleic Acids Research, 2010, Vol. 38, Database issue D318–D319.

NOVEL APPROACHES TO RNA SEQUENCING

Mazur A., Artemov A.V.
ZAO "Genoanalytica", Moscow, Russia

Motivation and Aim: RNA-seq analysis became the most powerful tool to explore disease-related phenomena. On one hand, transcription, unlike genotype, describes a phenotype of a cell. On the other hand, compared to different approaches of massive phenotype analyses, e.g. proteomics, RNA-seq has become much more comprehensive, sensitive and unbiased after the development of modern sequencing methods.

In this work we performed a comprehensive RNA-seq analysis of a skin disease. We explore not only disease-associated gene expression changes, but also shifts in alternative splicing patterns (including appearance of novel splicing isoforms) and changes in the expression within intergenic genome regions.

In order to reduce the influence of between-individual variation, we took two samples of the same tissue for every individual: one affected by the disease and the other not affected. 4 individuals participated in the study giving the total of 8 samples analyzed.

After the removal of ribosomal RNA, a paired-end sequencing (50bp from one side of a fragment and 35bp from the other) of the total RNA was performed on a SOLiD4 machine. The obtained colorspace reads were mapped to the human genome (hg19).

Methods and algorithms

On the first step, we performed a differential gene expression analysis with in-house R scripts utilizing 'edgeR' package. Taken the number of reads mapped to all exons of all annotated genes in all the samples analyzed, we estimated a p-value indicating if a gene's expression level differs between the samples affected and not affected by the explored disease.

On the next step we compared splicing patterns between affected and not-affected tissue samples. Paired-end reads were mapped with respect to potential exon-exon junctions by tophat software. Particularly, two reads of a pair were allowed to be mapped to different positions of genome and moreover every read were split into two parts and each of those was mapped separately. This mapping was further used to search for the splicing events annotated in Gencode V7 database which significantly differ between the samples of skin affected and not affected by the disease. We also performed *de novo* splicing isoforms assembly (i.e. without taking any reference exon-intron annotation into account) to find novel splicing events.

On the third step we explored intergenic expression in the given samples. We first defined clusters (i.e., genomic regions) of RNA reads expressed in at least one sample. The mappings for all samples were merged together. Several approaches of clustering were implemented: peak calling analogous to the one performed while ChIP-seq data processing; defining a cluster by the minimum number of reads it contains and maximal gap length between adjacent reads in one cluster; defining a cluster by the minimal number of reads and the maximal length of a cluster. The third definition with minimal number of reads = 30 and maximal length = 500 bp appeared to be the best in terms of presence of the reads within a cluster in more than one sample. Next, for every cluster we quantified its presence in every sample by counting the reads mapped to the cluster. These data were then used for a differential expression analysis similar to the gene expression analysis described earlier.

Results: Within the top 100 significant genes we found over-represented Gene Ontology (GO) categories which were consistent with the current understanding of the studied disease. Among differentially spliced genes a particular category of receptors was found. The differentially expressed regions will be further annotated with respect to known genomic features.

Conclusions: RNA-seq proved to be a source of biologically relevant information in the study of a skin disease. Extraction of the splicing and intergenic expression patterns provides meaningful results in addition to the results of gene expression analysis.

MOLECULAR DYNAMICS SIMULATION OF NIP7 PROTEINS FROM HYPERTHERMOPHILIC ARCHAEA AT HIGH TEMPERATURE AND PRESSURE

Medvedev K.E.*[1], Afonnikov D.A.[1, 2], Vorobjev Y.N. [3]
[1] *Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia;*
[2] *Novosibirsk State University, Novosibirsk, Russia;*
[3] *Institute of Chemical Biology and Fundamental Medicine, SB RAS, Novosibirsk, Russia*
e-mail: kirill-medvedev@yandex.ru
** Corresponding author*

Key words: *Nip7 protein, high pressure, molecular dynamic simulation, specificity determining positions, water-protein interaction*

Motivation and Aim: 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 (0.1-300 MPa) and temperatures (300, 373K). We perform search for Specificity Determining Positions (SDPs) in homologous archaeal proteins with respect to the organism's habitat depth (deep/shallow water) and compare results with molecular dynamics simulation analysis.

Methods and Algorithms: The structure of P.furiosus protein was obtained by homology modeling using Nip7 from P.abyssi as template (PDBID 2p38). MD simulations and structure analysis were performed using GROMACS [1]. Search for SDPs was performed using SDPPred [2] and GroupSim [3] software.

Results: It is shown that the structure of Nip7 N-terminal domain is more stable and has smaller structure fluctuations than C-terminal (RNA-binding) domain. We demonstrated that protein from P.abyssi is more stable under extreme conditions than from P.furiosus. P.abyssi protein has larger polar solvent accessible surface area in comparison with P.furiosus protein. Most of detected SDPs in Nip7 homologs display significant changes in side chain polarity between deep/shallow water organisms.

Conclusion: In general, these data demonstrate the importance of water-protein interactions for the protein stability under high pressure.

Acknowledgements: The work supported by RFBR (11-04-01771-a), SB RAS (projects 130, 39, 93), RAS (project 6.8 and program 28), State contract 82/201 and Scientific school 5278.2012.4.

References:

1. D.Van der Spoel et al. (2005) GROMACS: Fast, Flexible and Free. J. Comp. Chem. 26:1701-1718.
2. Kalinina OV, Mironov AA, Gelfand MS, Rakhmaninova AB. (2004) Automated selection of positions determining functional specificity of proteins by comparative analysis of orthologous groups in protein families. Protein Sci 13(2): 443-56
3. Capra JA and Singh M. (2008) Characterization and Prediction of Residues Determining Protein Functional Specificity. Bioinformatics, 24(13): 1473-1480.

INFLUENCES OF PROTEIN FUNCTIONAL SITES ENCODING FEATURES ON PROTEIN EVOLUTION IN EUKARYOTA

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

Key words: *protein functional sites, exon structure, protein evolution*

Motivation and Aim: Search of interrelationships between the structural–functional protein organization and exon structure of encoding gene provides insights into issues concerned with the function, origin and evolution of genes and proteins. The functions of proteins and their domains are defined mostly by functional sites. The relation of the exon–intron structure of the gene to the protein functional sites has been little studied. In the same time it could be useful in detection of exons involved in shuffling in an evolutionary perspective and helpful in rational design of novel proteins composed of fragments encoding individual exons from distinct genes.

Methods and Algorithms: To analyze protein functional sites encoding features we used earlier constructed computer system SitEx that includes database mapping the protein functional sites (FS) positions on the exon-intron structure of encoding gene and BLAST and 3DExonScan search.

Results: Testing SitEx system on proteins that possess known PDB structures reveal that exons encoding FS are more likely to be found and are more conservative than those not encoding. Also we have shown that the FS discontinuity through exon structure is significantly less than expected by chance and the protein FS tends to be encoded by single or neighboring exons. We analyzed codon usage and discovered that the exons coding FS amino acids on the 5`-end possess the less optimal codon usage. We found that the frequency of codons encoding FS on the exon border in phases 1 and 2 is significantly greater than it is for the others. It could be the result of the unification of the genes that encode the single FS. We analyzed SNP occurrence in FS in nearest area and found that the deleterious SNP are rare in FS, but SNP frequency is higher than in protein domains. The major physical factor for occurrence of non-synonymous coding SNP is amino acid hydrophobicity.

Conclusion: Protein functional site is highly conserved and fundamental identity that forms the evolutionary unit (for example domain). Its conservation arose from coding exon structure and is controlled by translation process. In the same time functional site amino acids are still available for evolutionary selection.

Acknowledgements (if necessary): Ministry of Science & Education (grant numbers 14.740.11.0001, 07.514.11.4003, in part); Interdisciplinary Integrative Project 35 of SB RAS (94, 111, 119, in part); Russian Foundation for Basic Research (grant no. 11-04-92712, in part); FP7: EU-FP7 SYSPATHO No. 260429; Program of RAS (A.II.5, A.II.6, B.21, B.26, in part).

CLUSTERIZATION OF GENE EXPRESSION PROFILES OF HUMAN ASTROCYTIC GLIOMAS ON SELF-ORGANIZING MAPS

Mekler A.A.*[1], Schwarz D.R.[1], Dmitrenko V.V.[2], Rymar V.I.[2], Iershov A.V. [2], Kavsan V.M. [2]

[1] *The Bonch-Bruevich Saint-Petersburg State University of Telecommunications;*
[2] *Institute of Molecular Biology and Genetics of NASU, Kiev Ukraine*
e-mail: mekler@yandex.ru
* Corresponding author

Key words: *gene expression, artificial intelligence, self-organizing maps, gliomas*

Motivation and Aim: Glial tumors are divided into astrocytomas and oligodendrogliomas. Glioblastoma is the most aggressive form of the gliomas. Malignant progression of the gliomas is accompanied by the accumulation of gene expression changes as compared to human normal brain. In tumors of various types, deviations of gene expression levels are different. The study represents an attempt to build a system for classification of astrocytic gliomas of different malignancy grades by the genes expression profiles using artificial intelligence systems.

Methods and Algorithms: In the analysis of multidimensional features a special emphasis is on data visualization that is such a representation of multivariate data on the two-dimensional plane, which at least qualitatively reflects the main regularities. This problem may be solved using the Kohonen self-organizing maps (SOM). SOM is a neural network for automatic clustering (unsupervised), realized in the form of a lattice of neurons. Each neuron (node lattice) corresponds to a vector whose dimension is equal to the dimension of the feature space. After training of this network, the most similar vectors that characterize the training set are displayed on the map near each other. In the present study, results of the clusterization of astrocytic gliomas on the basis of the expression of 20 genes are presented.

Results: SOM has been trained (using vectors of the 20 genes) by 5 training sets corresponding to 5 tissue kinds (classes) – healthy tissue, and 4 grades of astrocytomas (by the WHO classification). Sample sizes – 74, 45, 17, 93 and 224 samples for normal brain and tumors of the malignancy grades I, II, III and IV, respectively. Obtained map showed that genes expression profiles of some of the classes are well clustered on it and some are mixed.

Conclusion: From our point of view, using of the extended gene set for gene expression profile classification and correct gene selection may improve the clustering quality. This will lead to the possibility of tumor grade evaluation based on artificial intelligence systems (perceptron, etc.) [1]. In addition, there is a prospect of using these methods for more differentiated diagnosis.

Acknowledgements: Study was supported by the RFBR grant # 12-04-90434-Укр_а and NASU grant 07-0412.

References:

1. A.A. Mekler, D.R. Schwarz. (2010) Selection of variables for the reliable classification, In: XIII National Conference “Neuroinformatics-2011”: collection of scientific works in 3 volumes, vol. 1, 136-143.

ON METRIC PROPERTIES OF EVOLUTIONARY DISTANCES

Melchakova M.A.*[1], Efimov V.M.[2]
[1] *Novosibirsk State University, Novosibirsk, Russia;*
[2] *Institute of Cytology and Genetics, Novosibirsk, Russia*
e-mail: mariya.melchakova@gmail.com
** Corresponding author*

Key words: *evolutionary distance, Euclidean embedding, Kimura*

Motivation and Aim: We study the metric properties of evolutionary distances (p-distance, Jukes-Cantor, Kimura), which may not satisfy the triangle inequality. These metrics are defined on the sets of nucleotide sequences, and we are interested in possibility of embedding of these metrics into an Euclidean space. Such a representation can provide some information about evolutionary and structural relationships between these sequences.

Methods and Algorithms: Principal coordinates analysis was used to obtain spatial representation.

Results: It was shown that the Jukes-Cantor distance does not satisfy the triangle inequality, but it can be considered as a square of Euclidean distance. By analogy with the Jukes-Cantor distance, we have introduced (P,Q)-distance, which is also a square of Euclidean distance. We have considered the Kimura distance matrix of the *Adh* gene nucleotide sequences of drosophilid species [1] and empirically found that the square root transforms this distance matrix to that of the Euclidean distance. Results of application of principal coordinates spatial representation to this matrix correspond to the phylogenetic tree constructed in [1].

Conclusion: A method of transformation of the p-distances and the Jukes-Cantor distance to the Euclidean one is presented. This makes possible to use a set of algorithms of identification of structural relationships, for example the principal coordinates analysis. It has been suggested that Kimura distance can transformed to an Euclidean distance in a similar way, what was indirectly confirmed by the results of the paper [1].

References:

1. C.A.M. Russo, N. Takezaki et al. (1995) Molecular phylogeny and divergence times of drosophilid species, *Mol. Biol. Evol*. **12:** 391–404.

GENOME-WIDE ASSOCIATION STUDY OF CARDIOVASCULAR DISEASE RISK FACTORS IN THE MOSCOW STUDY OF THE WESTERN DISTRICT

Meshkov A.N.*[1], Khasanova Z.B., Konovalova N.V., Kotkina T.I., Sergienko I.V., Karpov Iu.A., Kukharchuk V.V., Boytsov S.A.[1]
Russian Cardiology Research Center, Moscow, Russia;
[1]*National Research Center for Preventive Medicine Moscow, Russia*
e-mail: meshkov@cardio.ru
** Corresponding author*

Motivation and Aim: Cardiovascular Disease is the leading cause of mortality in the Russian Federation. To identify common genetic polymorphisms associated with cardiovascular disease risk factors - total cholesterol, LDL- and HDL-cholesterol, triglycerides, lipoprotein(a), modified LDL-level, hypertension, smoking, and type-2 diabetes, we performed a genome-wide association study (GWAS) in 1,200 patients (366 male/834 female) from Moscow study of the Western District cohorts.

Methods and Algorithms: We genotyped 1,200 samples using the Illumina Cardio-Metabo BeadChips for a total of 196725 SNPs passing QC and allele frequency filters. Genotypes were called using a clustering algorithm in Illumina's BeadStudio software.

Results: We report 2 SNP associated with triglycerides level rs7259004 (p=9.623e-08) and chr19:50124992 (p=2.834e-07), 1 SNP associated with LDL-cholesterol level chr19:50103919 (p=6.654e-09), one cluster of 67 SNP located on chromosome 6 associated with lipoprotein(a) and one cluster of 14 SNP located on chromosome 2 associated with modified LDL level.

Conclusion: With the GWAS in the Moscow study of the Western District we confirm the previously found genetic associations with LDL-cholesterol levels, triglycerides levels, lipoprotein(a) levels and levels of modified LDL.

Acknowledgements: It is supported by Moscow City Government (#8/3-280н-10 and #8/3-281н-10).

EPIGENOMICS OF NUCLEOLAR DOMINANCE

Michalak P.
Virginia Bioinformatics Institute
Virginia Tech, Blacksburg, Virginia, USA
e-mail: pawel@vbi.vt.edu

Key words: *epigenetics, genomics, nucleolar dominance, rRNA, Xenopus*

Motivation and Aim. *Xenopus* are the only known vertebrates to tolerate aneuploidy, thus providing a unique model for successful management of changes in chromosomal numbers and competing genomes. *Xenopus* are also characterized by nucleolar dominance, a phenomenon by which cells of interspecies hybrids express the 45S rRNA precursor from the chromosomes of a single parental species only. This suppression results in a distinct reduction in the number of nucleoli formed in hybrid nuclei. Nucleolar dominance is one of the most dramatic but largely unexplored epigenetic reconfigurations in a cell. Analysis of the hybrid transcriptomes showed an asymmetric pattern of expression in relation to parental species, with genes from one of the parental species overall suppressed, suggesting that the genomic dominance extends beyond rRNA genes. The goal of the research is to characterize the mechanism of nucleolar dominance and asymmetric genome silencing using new genomic, epigenetic, and bioinformatics tools.

Methods and Algorithms. Next-gen sequencing methods (Ion Torrent, Illumina, and PacBio) were used for global expression, small RNA profiling and DNA methylation assays. Primary cells were treated with a number of epigenetic chemical inhibitors selectively targeting DNA methylation and histon modification enzymes.

Results. A combination of fluorescent staining and next-gen sequence profiling revealed nucleolar dominance in F1 hybrids between *Xenopus laevis* and *X. muelleri*, consistent with the overall pattern of gene misexpression. We also observed increased levels of DNA methylation in hybrids relative to parental species. Using a combination of chemical inhibitors, we partly derepressed nucleoli formation in hybrid cells.

Conclusion . New genomic and epigenetic approaches give valuable insights into nucleolar dominance. Nucleolar dominance provides a fascinating model for epigenetic control of large genomic segments.

Availability: Materials are available on request from the author.

LARGE-SCALE AMPLICON TARGETING MASSIVE PARALLEL RE-SEQUENCING REVEALS NOVEL VARIANTS IN ALZHEIMER'S DISEASE GENES

Mikhaylichenko O.A.[1, 2], Goltsov A.Y.[1, 2, 3], Gusev F.E.[1], Reshetov D.A.[1, 2], Tyazhelova T.V.[1], Andreeva T.A.[1,3], Kaljina N.R.[1,3], Grigorenko A.P.[1,3,4], Rogaev E.I.*[1,2,3,4]

[1] Vavilov Institute of General Genetics, Moscow; [2] Lomonosov Moscow State University; [3] Research Center of Mental Health, Russian Academy of Medical Sciences, Moscow, Russia; [4] Brudnick Neuropsychiatric Research Institute, Department of Psychiatry, University of Massachusetts Medical School, 01604, MA, USA

e-mail: EVGENY.ROGAEV@umassmed.edu

** Corresponding author*

Key words: *multiplex PCR, targeted resequencing, SOLiD, Alzheimer's disease*

Motivation and Aim: Emerging next-generation sequencing (massive parallel sequencing, MPS) technologies may transform the field of personalized medicine. Currently, direct sequencing of individual genome or exomes (all protein encoding genes) is already a feasible task. The next challenging task is to develop a large-scale MPS analysis of targeted genes in an extra- large clinical or population cohort, which could not be performed by conventional sequencing approach. We developed here the time- and cost-effective experimental and bioinformatics methodology for MPS analysis of disease-related genes. To perform the proof-of-concept study we selected the presenilin (*PSEN1, PSEN2*) and Amyloid precursor protein (APP) genes bearing mutations in familial cases of early Alzheimer's disease (AD). The comprehensive analysis of these genes is required to elucidate their role in the most common "sporadic" AD. Thus, we employed the un-biased clinical cohort of AD patients regardless their familial history and age-of onset. In addition, several other genes interacting with functional pathway of presenilins and APP were included in the study. A pipeline for MPS analysis of pooled multiple gene amplicons, threshold-criteria for filtration of sequencing errors and ready-to-run solutions for variant predictions has been developed and tested in a large size clinical cohort sample.

Methods and Algorithms:(1) We developed methodology for "all exons" multiplex amplification in single reactions for encoding regions of *APP,* and γ-secretase complex genes (*PSEN1, PSEN2, PEN-2, APH1A, NICASTRIN*), risk factor of disease *APOE* and the cleaving APP enzyme *BACE1.* In total, the samples from 552 AD and control individuals were tested. (2) The optimal algorithms for pooling of individual genomic DNAs and bar-coded MPS libraries were generated. The bar-coded pooled amplicon sets were sequenced using SOLiD™ 4 MPS platform. The raw reads were aligned with BioScope software and processed with in-house created PrimerCut tool which we developed to discriminate between genomic and synthetic PCR primer sequences. (3) Aligned and processed reads were transformed into pileup format with SamTools (http://samtools.sourceforge.net/); SNP calling was based on predicted variation frequency in a pool of N individuals/n reads.

Results: We showed the feasibility of analysis of 80 exons (18696 bp overall length)/per individual for >500 individuals in a single SOLiD™ 4 run and identified novel mutations in AD genes. The proposed bioinformatics pipeline is a ready-to-use solution for pooled DNA targeted re-sequencing analysis for both the research and medical purposes.

Availability: The PrimerCut tool is available on request.

Acknowledgements: This work was supported by Grant № 16.512.11.2083 "Research and development on priority directions of scientific-technological complex of Russia for years 2007-2013".

COMBINED *IN SILICO/IN VIVO* ANALYSIS OF AUXIN MEDIATED MECHANISMS OF ROOT APICAL MERISTEM DEVELOPMENT

Mironova V.V.*[1], Omelyanchuk N.A.[1], Novoselova E.S.[1], Doroshkov A.V.[1], Kazantsev F.V.[1], Kochetov A.V.[1, 2], Mjolsness E.[3], Likhoshvai V.A.[1, 2]

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

[2] *Novosibirsk State University, Novosibirsk, Russia;*

[3] *Departments of Computer Science and Mathematics; Institute for Genomics and Bioinformatics, University of California, Irvine, USA*

** Corresponding author: kviki@bionet.nsc.ru*

Motivation and Aim: Root apical meristem (RAM) is the plant stem cell niche which provides for the formation and continuous development of the root. Auxin is the main regulator of RAM functioning and auxin maxima coincide with the sites of RAM initiation and maintenance. Auxin gradients are formed due to local auxin biosynthesis and polar auxin transport. The PIN family of auxin transporters plays a critical role in polar auxin transport, and two mechanisms of auxin maximum formation in the RAM based on PIN mediated auxin transport have been proposed to date: the reverse fountain and the reflected flow mechanisms.

Methods: We combined both mechanisms in *in silico* studies of auxin distribution in intact roots and roots cut into two pieces in the proximal meristem region. The 2D mechanisms model was adapted from the model of reflected flow mechanism [1] and calculated using MGSModeller [2]. Numerical calculations were performed on SB RAS supercomputer cluster [2].

In parallel, we performed the corresponding experiments *in vivo* using DR5-GFP *Arabidopsis* plants [3].

Results and conclusions: By numerical simulations we showed that the reverse fountain and the reflected flow mechanisms naturally cooperate for the RAM patterning and maintenance in the intact root. Regeneration of the RAM in the decapitated roots is provided by the reflected flow mechanism. In the excised root tips local auxin biosynthesis along or in cooperation with the reverse fountain enables RAM maintenance. The model was also extended describing (1) different auxin synthesis mechanisms [3]; (2) DR5:: GFP expression [4]; (3) auxin distribution in *pin* mutants; (4) root treatments by active substances. In numerical experiments we demonstrated the efficiency of a dual mechanism model in guiding biological experiments on RAM development. The dual mechanism model can be a powerful tool for study several different aspects of auxin function in root.

Acknowledgements: Numerical calculations were performed on supercomputer cluster of Shared Facility Center for Bioinformatics SB RAS. Microscopy was performed in the Shared Facility Center for Microscopic Analysis of Biological Objects SB RAS. The work is partially supported by the Dynasty Foundation grant for young biologists, RAS program A.II.6, Integration project SB RAS 80 and RFBR grants 10-01-00717-a, 11-04-01254-a. EM was supported by NIH grant R01GM086883.

References:

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

MATHEMATICAL MODELING LANGUAGES FOR MORPHODYNAMICS

Mjolsness E.
University of California, Irvine, USA
e-mail: emj@uci.edu

Key words: *Dynamical grammar, morphodynamics, modeling language*

Motivation and aims: Biological development is a rich subject for mathematical modeling, and for computer languages that can support such modeling. Such languages should express mathematical models of objects, such as molecules and cells, and of dynamical processes that change the object state.

Results and conclusion: If the objects are essentially pointlike then we can define the semantics of reaction or rule based models using operator algebras, unifying discrete stochastic events and ordinary or stochastic differential equations. If reaction/rules are aggregated together, time evolution operators are summed, so the semantics is "compositional". Since objects can have state it is possible to represent discrete graphs (such as molecular complexes or cell assemblages) and their dynamics. These facts underlie "Dynamical grammar" modeling languages for development and morphogenesis. Can one define more generally the compositional semantics of biological object models including continua (such as membranes and branches filaments) on the same footing as discrete graphs? And what would the algebra of allowed processes look like for such compositional objects? The goal in asking these questions is to create computer languages for morphodynamics that are compositional and expressive both in their objects and processes.

THE GENOME OF THE CTENOPHORE *PLEUROBRACHIA BACHEI*: NEW INSIGHTS INTO EVOLUTION OF METAZOA AND ORIGIN OF NERVOUS SYSTEMS

Moroz L.L.[1,9], Kohn A.[1], Grigorenko A.P.[2,3], Yu F.[1], Farmerie W.[1], Citarella M.[1], Tyazhelova T.V.[3], Reshetov D.A.[3], Bostwick C.[1], Winters G.[1], Dabe E.[1], Povolotskaya I.[4], Kocot K.[5], Halanych K.[5], Gusev F.E.[2], Kondrashov F.A.[4], Solovyev V.[6], Ross J.[1], Rubakhin S.[7], Romanova E.[7], Daily C.[7], Sweedler J.[7], Berezikov E.[8], Rogaev E.I.*[2,3]

[1] University of Florida, Gainesville, USA; [2] University of Massachusetts Medical School (UMMS), BNRI, USA; [3] Vavilov Institute of General Genetics, Center of Genomics, Moscow; [4] Centre for Genomic Regulation, Barcelona, Spain; [5] Auburn University, USA; [6] University of London, London, UK; [7] University of Illinois, USA; [8] Hubrecht Institute, Netherlands; [9] University of Washington, Seattle and Friday Harbor, USA

** Correspondence: to L.Moroz. and to E.Rogaev.*

Key words: *ctenophore, Pleurobrachia, whole-genome*

Motivation and Aim: Our understanding of the animal origin is incomplete because of limited data from basal Metazoa. The phylum *Ctenophora* is one of the earliest animal lineages, but the phylogenic relationship of this group to Bilateral and non-Bilateral branches of metazoan animals remained controversial.

Methods and Algorithms: We have performed the complete genome sequencing for *Pleurobrachia bachei*. It has one of the most compact genomes within Ctenophore group. These holoplanktonic predators have sophisticated ciliated locomotion and well-recognized nervous and "true" mesoderm-derived muscular systems. Using 454/Roche, Illumina GA and HiSeq2000 sequencing platforms we achieved at least ~1,000x coverage of the genome. We performed RNA-seq profiling from major tissues to validated the genome assembly: 96% of predicted gene models are supported by transcriptome data.

Results: Our phylogenomic analysis demonstrates the most basal phylogenetic position of Ctenophores within the animal tree. This hypothesis is further supported by comparative analysis of selected gene families (including apparent lack of HOX genes and certain genes for canonical miRNA machinery). However, there is no evidence for elements required for serotonin, melatonin and histamine synthesis or conversion of phenylalanine to tyrosine. Our data of functional pathways reveal the Wnt, Notch classical MAPK, JNK and p38 MAPK signaling systems in *Pleurobrachia*. The genes for canonical proteolysis pathways including, e.g., intramembrane aspartic proteases, IMPAS/SPP and presenilin homologs, have also been identified. Apparently, there are no Toll-like receptors and cell-adhesion molecules, which are essential components in immunity. It is of great interest that our preliminary biochemical analysis and *in situ* hybridization showed no evidence for canonical neurotransmitters (e.g., dopamine and serotonin) in *Pleurobrachia*, implying that completely different neurosignaling molecules can be utilized in this branch of the Animal Kingdom. Our experimental data indicate that the nervous system in ctenophores is one of the most distinct in its morphological and molecular organization. Many "classical bilaterian neuron-specific" genes either not present or, if present, they are not expressed in neurons. Finally, we identified novel markers for ctenophores neurons. These data suggest that at least some of the ctenophore neural systems were evolved independently from those in other animals.

Acknowledgements. Supported by NSF, NIH, McKnight Brain Research Foundation, University of Florida, University of Washington and Friday Harbor, University of Massachusetts MS and Rostock group. Evgeny.Rogaev@umassmed.edu

PHYLOGENOMIC ANALYSIS OF DIATOM CHLOROPLAST GENOMES

Morozov A. A., Galachyants Y.P.*
Limnological Institute, SB RAS, Irkutsk, Russia
e-mail: yuri.galachyants@lin.irk.ru
** Corresponding author*

Key words: *diatoms, phylogeny, chloroplast, genome rearrangements, acpp*

Motivation and aim: Phylogeny of diatoms is still controversial, and phylogenomic analysis may be able to resolve it more reliably than analysis of single genes and morphological traits. Here we present an analysis of genome rearrangements in chloroplast genomes of diatoms – extremely diversified and largely unexplored group of unicellular eukaryotic algae that belong to heterokonts.

Methods and algorithms: Rearrangement analysis was carried out using MGR package [1]. RAxML 7.0.4 [2] was used for phylogenetic reconstruction.

Results: We analyzed the gene order in eight chloroplast genomes of diatoms and dinoflagellates bearing the diatom-derived plastids. The obtained tree topology was congruent with the generally accepted branching order within diatoms. Events of the gene acquirement and loss were revealed by comparison of the chloroplast genomes of diatoms [3] and mapped to corresponding branches of this tree. These events could be classified into three groups:

1) Gene loss driven by endosymbiotic gene transfer to the host nuclear genome. Lost genes are absent only in one or two genomes, like *psb*I in *Fistulifera* sp. or *pet*F in *Thalassiosira oceanica.* On the other hand, the presence of *tsf* in *Phaeodactylum tricornutum* and *Fistulifera* sp. chloroplast genomes is assumed to be the gene transfer in progress [4]. All genes which have no evidence of other scenarios were also assigned to this type, even in the absence of nuclear data.

2) Gene acquirement from chloroplast plasmids. *Ser*C1 and *tyr*C in *Kryptoreridinium foliaceum*, *ser*C2 in *K. foliaceum* and *Fistulifera* sp. JPCC DA0580, and several unidentified ORFs localized near *ser*C2 gene are assumed to be originated in plastid genomes of these two species from plasmid sequences.

3) Horizontal gene transfer to a host nuclear genome accompanied by gene loss from the chloroplast genome. This scenario has been shown to occur independently for acpp gene in several diatom species.

Conclusion: Analysis of the order of genes was shown to be applicable for inferring the diatom phylogeny and additional sequencing effort is required for reconstructing the reliable and representative trees.

Acknowledgements: This work was supported by the Program of the Presidium of the Russian Academy of Sciences "Molecular and Cell Biology" (project # 6.3).

References:

1. G. Bourque, P.A. Pevzner. (2002). Genome-scale evolution: Reconstructing gene orders in the ancestral species, Genome Res., 12: 26-36.
2. A. Stamatakis. (2006) RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models, Bioinformatics, 22: 2688-2690.
3. Y.P. Galachyants *et al*. (2012). Complete chloroplast genome sequence of freshwater araphid pennate diatom alga *Synedra acus* from Lake Baikal, Int. J. Biol., 4: 27-35.
4. M.-P. Oudot-LeSecq *et al*. (2006). Chloroplast genomes of diatoms *Phaeodactylum tricornutum* and *Thalassiosira pseudonana*: comparison with other plastid genomes of red lineage, Mol. Genet. Genomics, 277: 427-439.

MASS-SPECTROMETRIC MEASUREMENT OF LEVELS AND ENZYMATIC ACTIVITY OF CYTOCHROMES P450

Moskalyova N., Zgoda V.G.*, Tikhonova O., Novikova S., Kopylov A., Archakov A.I.
Orekhovich Institute of Biomedical Chemistry, Russian Academy of Medical Sciences
e-mail: vic@ibmh.msk.su
** Corresponding author*

Key words: *cytochrome P450 (CYP), multiple reaction monitoring (MRM), mass-spectrometry, enzymatic activity, drug oxidation*

Motivation and Aim: Personalized medicine requirements dictate the possibility to study metabolism of several drugs in individual patient. It is well known that cytochromes P450 (CYP) play a crucial role in oxidation of most medicines. Thus, a need is in new multiplied methods to measure a signature of levels of these enzymes and their activities.

Methods and Algorithms: Levels of several members of mammalian CYP subfamilies 1A, 3A, 1E, 2C, 2D were measured by multiple reaction monitoring in triple quadrupole mass spectrometer. Method was developed and validated using samples of murine liver microsomes taken from intact controls and mice induced by phenobarbital and methylcholanthrene xenobiotics.

Results: The method allowed reliable measurements of levels of several CYP isoforms without use of isotopic molecular mass labels and specific derivatizing agents. The results correlated with values of enzymatic activity determined using marker substrates specific for CYP isoforms of interest.

Conclusion: The new method to measure cytochrome P450 signature by state-of-the-art mass-spectrometry is developed which may be easily translated to human healthcare.

Acknowledgement: The work was funded by Russian Academy of Medical Sciences.

COMPUTATIONAL ANALYSIS OF NON-BONDED INTERACTIONS BETWEEN ATOMS OF PROTEIN AND MEDIUM REPLACING RMSD METRIC

Mukha D.V.*, Usanov S.A.
Institute of Bioorganic Chemistry NAS of Belarus, Minsk, Belarus
e-mail: dmitry.mukha@iboch.bas-net.by
** Corresponding author*

Key words: *molecular dynamics, RMSD, molecular interactions*

Motivation and Aim: Water medium plays an important role in processes of proteins' functioning in biological systems. Effects of entropy changing due to condensation of polypeptide chain and formation of secondary structure, mainly, drive a process of folding [1]. Water molecules located in cavities and pockets of proteins are important for native structure maintaining and molecular interactions with ligands [2].

Methods and Algorithms: The main idea of the approach is to account the ratio between time of residence for diffusing molecules in the equal volume sites standing near and far from biopolymer. This ratio is determined by means of molecular dynamics simulation. Values of time periods are averaged along time coordinate. Thus, self-diffusion properties for water molecules can be transformed into energy of binding estimate. For spatial data storage Gaussian Cube (Gaussian, Inc.) file format was adapted.

Results: We developed a novel approach to analysis of non-bonded interactions between atoms of protein and medium. This approach is free from drawbacks of RMSD, such as dependency on time window for analysis. This is due to fact that the novel approach does not operate with numbered atoms, but instead of this, with spatial cells in coordinate system aligned to biopolymer structure. The approach has been tested on a set of 45 precise crystallographic structures from PDB. The analysis showed that the approach has a capacity to correct a reconstruction of protein-ligand interaction allowing to account tightly bound water molecules during docking procedure.

Conclusion: Thus, we developed the new approach for studying of diffusion and binding of molecules from a medium surrounding a protein. The approach is based on molecular dynamics simulation. Uncovering the essential role of water in catalysis of most reactions having biological importance the approach presented provides a way of studying molecular recognition mechanisms and principles of enzymatic functioning.

Availability: A software tool written on JAVA and implementing the approach described is free for non-commercial use in academic institutions.

References:

1. Yasuda, S., et al., *Effects of side-chain packing on the formation of secondary structures in protein folding.* J Chem Phys, 2010. **132**(6): p. 65105.
2. Yin, H., et al., *Water in the polar and nonpolar cavities of the protein interleukin-1beta.* J Phys Chem B, 2010. **114**(49): p. 16290-7.

HAPLOID EVOLUTIONARY CONSTRUCTOR: PARALLELIZATION AND HIGH PERFORMANCE SIMULATIONS OF PROKARYOTIC COMMUNITIES EVOLUTION

Mustafin Z.S., 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*

Motivation and Aim: Haploid evolutionary constructor (HEC) is a software package which is used for the simulation of evolution of prokaryotic communities. HEC models describe several levels of biological organization (genetic, metabolic, population, ecological) [1], which cauzes simulations to take much computational time. For example, some of complex HEC models which describe evolution of high-diversity microbial communities may take dozens of hours when computed on the ordinary PC. This study was aimed to accelerate HEC simulations by developing a high performance version of HEC and using a parallel cluster computing (http://evol-constructor.bionet.nsc.ru).

Methods and Algorithms: We have developed the optimized algorithm for calculation of the polymorphic population growth, which was found to be the most time-consuming stage in HEC simulation process. A sufficient part of this stage was spending on the generation of all possible alleles combinations in a population. In the original HEC algorithm for such generation was used the simple recursion. We did recursion unwind, which itself gave us about 25 % of performance increasing. Hereafter we have developed the parallelized version of our algorithm using the MPI (Message Passing Interface) technology for high performance calculations. The tests and benchmarks have been performing on the clusters of the Center for collective usage "Bioinformatics" of the SB RAS (http://bioinformatics.bionet.nsc.ru/), and the Siberian Super Computer Center (http://www2.sscc.ru/).

Results: The overall speedup of the parallelized version of HEC runtime was found to be near-linear in some appropriate conditions. In order to achieve such conditions we should simulate a "well-posed model", which contains high-diversity polymorphic populations and not contains low-diversity polymorphic or monomorphic ones. For example, our first "well-posed model" contained the polymorphic population of very high genetic diversity (10^8 possible allelic combinations). The simulation of only 100 generations of this model requires about 763 minutes using the original HEC program, about 572 minutes using the modified HEC program (recursion unwind) and only 9 minutes using the MPI version (8 nodes, 64 MPI processes).

Conclusion: The developed algorithm allows a user to increase the simulation process of complex high-diversity prokaryotic communities significantly. We hope it will conduce getting novel results in theoretical evolutionary biology.

Acknowledgements: The work was supported by the RFBR grants 10-04-01310-a, 12-07-00671-a, Interdisciplinary integration projects of SB RAS № 47, 130.

References:

1. S.A. Lashin et al. (2012) Computer modeling of genome complexity variation trends in prokaryotic communities under varying habitat conditions, *Ecol. modelling,* **224:** 124-129.

EVOLUTION OF THE α-L-RHAMNOSIDASES: HISTORY OF THE LATERAL GENE TRANSFERS AND THE GENE DUPLICATIONS

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

Key words: *glycoside hydrolase, α-L-rhamnosidase, protein evolution, lateral gene transfer, molecular phylogeny, paralog, GH78, GH106, CAZy, TIM-barrel, PSI-BLAST*

Motivation and Aim: The α-L-rhamnosidase (EC 3.2.1.40) is a widespread and industrially important enzyme, which catalyzes the hydrolysis of the terminal, non-reducing α-L-rhamnose residues in various carbohydrates and their derivatives. The catalytic domains of the majority of currently known enzymes of this group belong to the GH78 and GH106 glycoside hydrolase families. According to the CAZy database ({{http://www.cazy.org/}}), these families consist of 428 and 80 proteins, respectively. The former includes proteins with $(\alpha/\alpha)_6$-barrel type of the three-dimensional structure (PDB, 2OKX and 3CIH), while the structure of the latter is still unknown. Relationships within the GH78 and GH106 families are unclear and became the purpose of the work, as well as evolutionary connections with other protein families.

Methods and Algorithms: Protein sequences were retrieved from the NCBI database. Multiple sequence alignments of 358 and 343 domains from GH78 and GH106 families, respectively, were made in BioEdit program. The phylogenetic trees were built using Neighbor-Joining and Maximum Parsimony programs of PHYLIP package. Interfamily relationships were established using PSI Protein Classifier. The program summarizes results of both successive and independent PSI-BLAST searches. Several most divergent representatives from the GH78 and GH106 families were used as queries.

Results and Conclusion: More than 1700 non-identical protein sequences of GH78 and GH106 domains have been revealed using the blast algorithm. They represent several phyla of Bacteria, as well as some Archaea and Eukaryota (mainly fungi). Proteins from the phyla Acidobacteria, Bacteroidetes, and Verrucomicrobia were overrepresented compared to the phyla Actinobacteria, Firmicutes, and Proteobacteria. Phylogenetic analysis of GH78 and GH106 families suggested multiple events of lateral gene transfer, as well as the gene duplications and eliminations. In particular, genes of the majority of 22 acidobacterial hypothetical α-L-rhamnosidases belonging to 13 distinct clusters evolved as the result of several independent gene transfers from the Bacteroidetes. *Clostridium methylpentosum* (GenBank, ACEC00000000.1) encodes at least 83 proteins containing GH78 or GH106 domains, while the majority of other sequenced *Clostridium* genomes do not contain them at all.

Iterative screening of the protein database allowed us to reveal relationship of GH78 with several other families of glycoside hydrolases having the $(\alpha/\alpha)_6$-barrel type of the three-dimensional structure: GH15 (clan GH-L), GH37 (GH-G), GH63 (GH-G), GH65 (GH-L), GH92, GH94, GH95, GH100, and GH116. PSI-BLAST searches using GH106 domains as queries found their relationship with the $(\beta/\alpha)_8$-barrel domains belonging to GH2 (clan GH-A), GH5 (GH-A), GH13 (GH-H), GH35 (GH-A), GH39 (GH-A), GH42 (GH-A), and GHL33 families of glycoside hydrolases.

DEEP SEQUENCING CHRYSANTHEMUM microRNA ON DIFFERENT STAGES OF PLANT DEVELOPMENT

Nedoluzhko A.V. *[1], Pantiukh E.S.[2], Rastorguev S.M.[1], Gruzdeva N.M.[1], Shulga O.A.[3], Prokhortchouk E.B.[1, 3], Skryabin K.G.[1, 3]

[1] *National Research Centre "Kurchatov Institute" (NRC "Kurchatov Institute"), Moscow, Russia;*

[2] *Lomonosov Moscow State University (MSU), Moscow, Russia;*

[3] *Centre "Bioengineering" Russian Academy of Sciences, Moscow, Russia*

e-mail: nedoluzhko@gmail.com

** Corresponding author*

Key words: *MicroRNA, Chrysanthemum morifolium, Next Generation Sequencing*

Motivation and Aim: MicroRNAs (miRNAs) are small endogenous non-protein-coding RNA molecules (18-24 bp), which regulate different processes in cells. Plants, animals, fungus and even viruses have various miRNA families, but sometimes they are very conservative among evolutionally distinct taxa. To understand the molecular basis of microRNA-induced processes it's important to accurately quantify known miRNAs expression profiles as well as to discovery novel miRNAs.

Methods and Algorithms: In this study we have determined miRNA profiles of Chrysanthemum morifolium in three stages of plant development: premature stage, generative stage (budding) and generative stage (full flowering). The miRNA sequencing was performed using SOLiD 4.0. Sequenced miRNA data were mapped on plant mirBase (version 18.0) using Bowtie software.

Results: We have found a broad range of known microRNAs in Chrysanthemum genome and identified miRNAs that are differentially expressed between different stages of the plant development. For example, for the generative stage we have described widely presented miRNA families: miR396 (14680), miR167 (7431), miR166 (6967), miR162 (5059), miR159 (1965) and etc.

Conclusion: Furthermore, we have identified potentially novel microRNAs, which will be proved using other technics such as Sanger sequencing, Northern blotting and (or) bioinformatic analysis.

ON BIOMECHANICS OF ARABIDOPSIS EMBRYO AND INTERPRETATIONS FOR GEOMETRICAL FEATURES OF EMBRYO RECONSTRUCTIONS BASED ON CONFOCAL MICROSCOPY

Nikolaev S.V.*[1], Trubuil A.[2], Palauqui J.-C.[2], Kolchanov N.A. [1]
[1] *Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia;*
[2] *INRA, Paris, France*
e-mail: nikolaev@bionet.nsc.ru
** Corresponding author*

Key words: *plant embryo, development, confocal microscopy, biomechanics, shell mechanics*

Motivation and Aim Development of plant embryo beginning from the just early stages is not fully studied yet. For example, it is not known detaled description of dynamics of cellular composition of embryo, and the cells growth. Biomechanics appeared to play a significant role in growth and morphodynamics during embryo development. In this work we studied influence of intracellular osmotic pressure distribution and mechanical properties of the embryo cells on the embryo 3D geometry change.

Methods and Algorithms From a simplified biomechanics view the plant cell seems as an osmotic device included in a thin cell wall. So we can apply a formalism of mechanics of shells in this case. We tested elastic and elastoplastic models of material for the cell wall, and we looked for parameters in published papers. The Structural Mechanics Component of the COMSOL Multiphysics package was used to construct simplifyed 3D geometrical models of Arabidopsis embryo for 2-,4-, and 8-cells stages, and to perform calculations. Geometrical sizes for the 3D models were obtained from 3D reconstructions based on confocal scanning microscopy. For other part of mechanical modeling these 3D reconstructions of the embryo cells were used as embryo models with realistic geometry. We used STL format to export-import geometry between Avizo package and COMSOL; the former was used to 3D reconstructions from stack of confocal images.

Results 3D simplifyed geometrical models of early stages Arabidopsis embryodevelopment were constructed in shell representation. Different models of the cell wall materials were studyed. Calculations of stationary solutions for some different intracellular pressures were performed.

Conclusion Shell mechanics can be usefull approximation of the cell mechanics to model embryo cells to obtain some insight and understanding of features of observed confocal images and embryo reconstructions. It was demonstrated that some curved intercellular walls can be explained as a consequence of different intracellular pressures in the neighbour cells. Sometimes observed edges on embryo surface can be interpretated as an evidence of embryo shrinking under sample preparation.

Acknowledgements The work was partly supported by the RFBR grant 11-04-01748-a «Computer analysis and modeling of shoot apical meristem development», and the RFBR-INRA grant 11-04-91397 «Acquisition of bilateral symmetry in embryos of dicotyledonous plants».

PROF_PAT, THE DATABASE OF PROTEIN FAMILY PATTERNS – AN EFFECTIVE TOOL FOR SEQUENCES ANNOTATION

Nizolenko L.Ph.*, Bachinsky A.G.
FBIS NRC of Virology and Biotechnology "Vector", Koltsovo, Russia
e-mail: nizolenko@vector.nsc.ru
** Corresponding author*

Key words: *protein families; patterns; similarity search; data banks; amino acid sequences*

Motivation and Aim: In the case of distant proteins, the search for global similarity of complete sequences may fail to show positive result, because conservative blocks responsible for their special functions may prove to be relatively short and scattered all over the sequence. So, when analyzing novel protein sequences, in addition to routine searches of the primary data sources, it is essential to extend search strategies to include a range of "secondary" databases, representing protein families. The level of noise is lower, because comparison is usually made with patterns that represent conservative intervals of positions.

Methods and Algorithms: Prof_Pat is a sample of "secondary" database, created for the greatest possible number of proteins of the UniProt. Flexible program for fast searching is supplied. It can investigate individual amino acid sequence, as well as large set of them in one pass. Technology of construction of Prof_Pat is described in detail in [1].

Results: The main field of application of Prof_Pat is an annotation of the new amino acid sequences. Such as amino acid sequences, translated from complete genomes of micro organisms. An example of this type of the investigation is an analysis of the open reading frames of *M. tuberculosis* described in [2]. A brief summary of the study and results of the investigation of amino acid sequences of some other microorganisms are shown.

Microorganism name	Number of open reading frames	Results of comparison with Prof_Pat		
		Similarity not founded	Recognized with high significance level	New[a] similarity
M. tuberculosis	3924	2	2777	44
Bacillus subtilis	4105	4	3284	22
Salmonella typhi	4767	4	4578	16
Brucella abortus	1033	5	992	36

[a]This protein's functions were not predicted earlier.

Conclusion: Despite the existence of multiple databases and algorithms to search distant similarity of amino acid sequences, Prof_Pat provides innovative results and, therefore, serves as a useful tool in the study of new sequences.

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

References:

1. A.G. Bachinsky et al. (2000) PROF_PAT 1.3: updated database of patterns used to detect local similarities, *Bioinformatics*, 16: 358-366.
2. L.F. Nizolenko et al. (2005) Investigation of the amino acid sequences of open reading frames of the Mycobacterium tuberculosis complete genome with the use of the protein family pattern bank Prof_Pat, *Biofizika*, 50: 986-992.

UGENE ASSEMBLY BROWSER: A TOOL FOR NGS DATA VISUALIZATION

Novikov I.A.*, Fursov M.Y., Efremov I.E.
Novosibirsk Center of Information Technologies 'Unipro'
e-mail: inovikov@unipro.ru
** Corresponding author*

Key words: *bioinformatics, genome, next generation sequencing, visualization*

Motivation and Aim: Next generation sequencing methods has already become a very popular starting point of genome analysis. An important feature of NGS methods is the volume of produced data, which can exceed tens of gigabytes, thus presenting a challenge for developers of interactive visualization software. Users of such software typically need to get a complete overview of the whole data as well as to navigate between narrow regions of interest quickly. Moreover, it is usually needed to make further analysis of the data. Unipro UGENE [1] is a free bioinformatics suite incorporating popular algorithms and a convenient designer for organizing them into workflow. Assembly Browser is a quickly developing module of UGENE aimed for visualization and analysis of NGS data.

Methods and Algorithms: Assembly Browser stores aligned short reads in an embedded database. It makes possible to instantly access any region of short reads assembly/map without a need to load entire file into computer memory. Packing reads and computing overall coverage from scratch are time-consuming tasks, therefore they are done only once, during import of reads into the database. Reference nucleotide sequences are also imported into database when opened in UGENE instead of loading into memory entirely.

Results and Conclusion: We have developed an interactive software solution called Assembly Browser. It allows users to import data from SAM or BAM format and then get the overview with the most covered regions. Users can quickly navigate them even on computers with little memory. It is possible to view the reads, get their properties, highlight differences from reference, examine coverage graph and consensus sequence of a given region, as well as to export reads and consensus.

Availability: Assembly Browser is available as a part of UGENE suite. A ready to use version of UGENE as well as the complete source code is freely available under the terms of GPL license from the web site [2] or from the repositories of Ubuntu and Fedora Linux distributions. Binary packages are available for Linux, Windows and Mac OS X operating systems.

References:

1. Konstantin Okonechnikov, Olga Golosova, Mikhail Fursov, the UGENE team (2012) Unipro UGENE: a unified bioinformatics toolkit, *Bioinformatics,* doi: 10.1093/bioinformatics/bts091.
2. Unipro UGENE website: http://ugene.unipro.ru.

ROLE OF AUXIN DOSE-DEPENDENT CONTROL IN SPECIFICATION OF ROOT VASCULAR CELLS

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: esn@bionet.nsc.ru
** Corresponding author*

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

Motivation and Aim: Root central cylinder of vascular plants includes vasculature (xylem and phloem), parenchyma, cambium and pericycle. The vasculature of *Arabidopsis thaliana* has bilateral symmetry with two poles of xylem and two poles of phloem elements. The different poles are perpendicular to each other on a root cross-section. Specification of the root vasculature occurs in the meristematic zone. It was shown that PIN1 auxin carrier is expressed in the whole central cylinder of meristematic zone [1], while another carrier AUX1 – only in protophloem poles [2]. Expression of these auxin carriers are regulated by auxin in a dose-depended manner [1]. So far the definite mechanism of auxin action on PIN1 and AUX1 patterning in the root central cylinder has remain unknown. Here we study this issue by mathematical modeling methods.

Methods: The processes of PIN1 and AUX1 synthesis, diffusion and auxin polar transport into/out cells, auxin and carriers degradation are described using the law of mass action and in terms of generalized Hill functions [3]. Assembly of the model describing auxin distribution in 2D cell array was carried out in MGSModeller [4]. The model parameters were taken from [5] as well as adjusted by model calculations in Mathematica, MGSModeller и STEP+ software packages.

Results: The 2D model describing auxin distribution over root central cylinder cross-section was developed. The cell layout represents 37 cells arranged in a circle which imitates the root cross-section. The values of parameters were adjusted so that the pattern of auxin carrier's expression in the root central cylinder obtains bilateral symmetry features.

Conclusion: As a result of the model simulation we propose a hypothesis about the morphogenetic factors providing for bilateral symmetry in the root central cylinder of *A. thaliana*. They are (1) the non uniform auxin flow from the shoot to the root in the central cylinder and (2) regulation of carrier's expression by auxin in a dose-dependent manner.

Acknowledgements

The work is partially supported by the RAS program A.II.6, Integration project SB RAS 80, RFBR grants 10-01-00717-a,11-04-01254-a, the Dynasty Foundation grant for young biologists.

Reference

1. A Vieten et al., (2005), Development, 20:132.
2. R Swarup et al., (2001), Genes Dev.;15(20):2648-53.
3. V Likhoshvai, A Ratushny, (2007), J Bioinform Comput Biol., 5(2B):521-31.
4. F Kazancev et al., (2012) In this proceedings.
5. VV Mironova et al., (2010), BMC Systems Biology, 4:98.

THE ROLE OF MATURE microRNA NUCLEOTIDE CONTEXT IN THEIR FUNCTIONING

Omelyanchuk N.A.*, Ponomarenko P.M., Ponomarenko M.P.
Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
e-mail: nadya@bionet.nsc.ru
** Corresponding author*

Key words: *microRNA, nucleotide sequence, RISC, Argonaute proteins, affinity*

Motivation and Aim: MicroRNAs (miRNAs), endogenous small RNAs with a length of 20–24 nucleotides, regulate target mRNA stability and functioning by directing RNA induced silencing complex (RISC) to the complementary site in the mRNA. We have discovered before the correlation between the miRNA abundance in A*rabidopsis thaliana* organs and the presence of the tetranucleotide WRHW in the central part of the miRNA sequence and/or the tetranucleotide DRYD at its 3' end [1]. After than it was experimentally shown that Argonaute (Ago) proteins regulate microRNA stability and, by this way, increased microRNA abundance [2]. We use the data on affinity of human miRNAs for Ago2 and Ago3 proteins [3] to determine if this feature is also context dependent and what parts of the mature miRNA sequence (5' end, the center or 3' end) are responsible for this. When the human RISC contains Ago3 protein, the mRNA translation is inhibited; however, in the case of Ago2, the mRNA can be also cleaved in the center of mRNA/miRNA heteroduplex [4]. Due to the latter, the aim of our analyses was also to reveal whether Ago2 with its slicer function has a special affinity to miRNAs with particular nucleotide context.

Methods and Algorithms: In this work, we used the computer system ACTIVITY [5] developed to analyze the pairwise data of a "nucleotide sequence, quantitative value" type and applied earlier to different protein/DNA interactions.

Results: It has been found that the higher the abundance of YRHB tetranucleotides near the miRNA 3'end, the higher is the miRNA affinity for both Ago2 and Ago3 proteins. The miRNA affinity to Ago2 is higher than to Ago3 if this miRNA contains the RHHK tetranucleotides in the center of its sequence.

Conclusion: Along with the well known site for target mRNA recognition at the 5' end of the miRNA sequence, we revealed the role of the 3' end in miRNA affinity for Ago proteins, loading on which provides miRNA stabilization. Also we showed that the RHHK tetranucleotides near the miRNA center where Ago2 realizes its slicer function provides a special affinity of miRNAs to this protein. Because most of Arabidopsis miRNAs direct the RISC slicer acivity [6], it can explain, why their abundance is determined by special nucleotide context in both 3' end and the center of the miRNA sequence.

References:

1. M. Ponomarenko et al. (2008) *Doklady Biochemistry and Biophysics*, **420:** 150–154.
2. J. Winter, S. Diederichs. (2011) *RNA Biol*. **8:** 1149-1157.
3. A. Azuma-Mukai A. Et al. (2008) *Proc.Natl. Acad. Sci. USA*. 105: 7964–7969.
4. A. Lingel, E.Izaurralde (2004) *RNA*. **10:**1675-1679
5. M. Ponomarenko et al. (1997) J. Comput. Biol. **4:** 83–90.
6. N. Baumberger, D. Baulcombe (2005) Proc Natl Acad Sci U S A. **102:**11928-11933.

PLACE MAKES A SEQUENCE: THE INFLUENCE OF HIGH AND LOW COPY REPEATS ON THE ORIGIN AND FATE OF MICROSATELLITES IN VERTEBRATE GENOMES

Oparina N.Y.*[1, 2], Fridman M.[3], Kulakovskiy I.V.[3], Makeev V.J. [3]
[1] *Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow, Russia;*
[2] *Orekhovich Institute of Biomedical Chemistry of the Russian Academy of Medical Sciences, Moscow, Russia;*
[3] *Vavilov Institute of General Genetics, Russian Academy of Sciences, Moscow, Russia*
e-mail: oparina@gmail.com
** Corresponding author*

Motivation and aim The microsatellites are known to be characterized by high lability mainly consisting of copy number variation. Despite decades of active research little is known about the mechanisms of both microsatellites generation, their (in)stability and varying evolutionary lifetimes.

Methods and algorithms We have carried out the comparative analysis of 3-6 bp microsatellites in vertebrate genomes paying special attention to their flanking regions and their characteristics. Genomic distribution of such repeats and associated elements was studied as well as their sequence characteristics.

Results We have demonstrated that the most frequent sequences of the longer microsatellite monomer types were mostly associated with recent subfamilies of active SINEs and also with recent segmental duplications. The "hot" regions of microsatellite generation were identified. We have shown the prevalent role of recombination-related events, not solely polymerase slippage during replication in microsatellite generation. The shorter microsatellite monomer types were shown to be generated through heterogeneous mechanisms. We have also demonstrated the significant impact of recombination-related events onto further fate of microsatellite, including their conversion to similar repeat types.

Conclusion We have received the fruitful results explaining the similar features of genomic instability in various vertebrate genomes despite their differences especially in non-coding DNAs.

CYTOCHROME P450 SUPERFAMILY IN VERTEBRATES: EVOLUTIONARY PATHS OF XENOBIOTIC CYP450 AND ENVIRONMENTAL «LIFESTYLES»

Oparina N.Y.*[1, 2], Zharkova M.[2], Speranskaya A.S.[1], Veselovsky A.[2]
[1] *Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow, Russia;*
[2] *Orekhovich Institute of Biomedical Chemistry of the Russian Academy of Medical Sciences, Moscow, Russia*
e-mail: oparina@gmail.com
** Corresponding author*

Motivation and aim The cytochromes P450 (CYP450) are the one of the largest and the most studies protein superfamilies. However, due to their variant functions, namely the metabolism of endogenous small molecules and in xenobiotic compounds detoxification, different CYP450 families evolve with extremely different rates. Vertebrate genomes contain the same 19 CYP450 subfamilies while the number of encoded genes varies a lot.

Methods and algorithms We carried out the comparative analysis of stable and unstable CYP450 families using a wide range of molecular evolutionary metrics. We have designed the "lifestyle" classification of environmental and food features of vertebrate phyla.

Results The resulting scale was used for identification of similar features of CYP450 gene duplications and differentiation in evolutionary distant taxa characterized with similar "lifestyles".

Conclusion We have shown that the formalized "lifestyle" scale made it possible to detect undiscovered similarities in CYP450 unstable families evolution in vertebrates. Our results shed light on the mechanisms and driving forces of xenobiotic CYP450 differentiation.

PROPERTIES OF miR156a AND miR171a BINDING SITES IN PROTEIN-CODING SEQUENCE OF PLANT GENES

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

Key words: *microRNA, mRNA, gene, Arabidopsis thaliana, plant*

Motivation and Aim: About 70% of miRNA binding sites in plant genes are located in mRNA protein-coding sequence (CDS). Individual miRNAs and miRNA families are identical or slightly different in closely related and phylogenetically distant plant species. It is required to determine the properties of miRNA binding sites in CDS of one family genes in different organisms.

Methods and Algorithms: Genes nucleotide sequences of *Arabidopsis thaliana* and other plants were obtained from GenBank (http://www.ncbi.nlm.nih.gov). miRNAs nucleotide sequences were received from miRBase (http://www.mirbase.org). The free energy of miRNA:mRNA hybridization was calculated using RNAHybrid 2.1. Graphs of the variability of nucleotide and amino acid sequences were created by WebLogo program.

Results: There were established ath-miR156a and ath-miR171a binding sites in CDS of target genes of *A. thaliana* and other plants. Targets for ath-miR156a are genes of SPL family (squamosa-promoter binding protein-like), encoding DNA binding proteins and transcription factors (*SPL2, SPL6, SPL9, SPL10, SPL11, SPL13, SPL15*). ath-miR156a binding sites in CDS of these genes of *A. thaliana* contain fully homologous GUGCUCUCUCUCUUCUGUCA polynucleotide which encodes ALSLLS hexapeptide in SPL proteins. This polynucleotide is perfectly homologous in ath-miR156a binding sites in SPL gene families of *Arabidopsis lyrata*, *Oryza sativa*, *Populus trichocarpa, Physcomitrella patens*, *Ricinus communis*, *Sorghum bicolor*, *Vitis vinifera*, *Zea mays*. Consequently, ALSLLS hexapeptide is also conservative in SPL proteins of these plant species. It was found that targets for ath-miR171a are GRAS transcription factor gene family of *A. thaliana* (HAM1, HAM2, HAM3). Binding sites for ath-miR171a localized in CDS contain GAUAUUGGCGCGGCUCAAUCA polynucleotide which encodes ILARLN hexapeptide in corresponding proteins. ath-miR171a:mRNA interaction sites in orthologous genes of *A. lyrata*, *B. distachyon*, *Glycine max*, *Medicago truncatula*, *O. sativa*, *P. patens*, *R. communis*, *S. bicolor*, *Selaginella moellendorffii*, *V. vinifera*, *Z. mays* also contain conservative polynucleotide. Orthologous HAM1, HAM2, HAM3 proteins contain conservative ILARLN hexapeptide but amino acids located before and after the hexapeptide are variable.

Conclusion: The polynucleotides of ath-miR156a binding sites found in CDS of SPL gene family are conserved and encode ALSLLS hexapeptide in orthologous SPL proteins. The polynucleotides of ath-miR171a binding sites in CDS of GRAS transcription factor gene family are conserved and encode ILARLN hexapeptide.

Availability: Obtained data can be used to regulate plants gene expression with ath-miR156a and ath-miR171a.

COMPUTER ANALYSIS AND DATABASE PRESENTATION OF ANTISENSE TRANSCRIPTION ASSOCIATED WITH microRNA TARGETS IN PLANT GENOMES

Orlov Y.L.[1], Chen D.[2], Dobrovolskaya O.[1], Meng Y.[2], Chen L.[2], Afonnikov D.A.[1], Chen M.*[2]

[1] *Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia;*
[2] *Zhejiang University, Hangzhou, China*
e-mail: mchen@zju.edu.cn
** Corresponding author*

Key words: *transcription, plant genomes, miRNA, sequencing, databases, genomics*

Motivation and Aim: Long non-coding RNA and non-coding genes, including small interfering RNAs (siRNAs), are key components of gene expression in eukaryotes, forming a regulatory network. miRNAs are expressed through nucleolytic maturation of hairpin precursors transcribed by RNA Polymerase II or III. Such transcripts are involved in post-transcriptional gene regulation in plants, fungi and animals. miRNAs bind to target RNA transcripts and guide their cleavage (mostly for plants) or act to prevent translation. siRNAs act via a similar mechanism of cleavage of their target genes, but they also can direct genomic DNA methylation and chromatin remodeling. It is estimated that large fraction, up to 30% of all human genes also may be post-transcriptionally regulated by miRNAs. Integration of these data in specialized databases is challenging problem of computer genomics.

Methods and Algorithms: To meet issues of statistical analysis of plant genome sequencing data we developed set of computer programs to define antisense transcripts and miRNA genes based on available sequencing data. We had search for homological regulatory regions in model plant genome organisms.

Results: We have analyzed data from PlantNATsDB (Plant Natural Antisense Transcripts DataBase) which is a platform for annotating and discovering NATs (Natural Antisense Transcripts) by integrating various data sources [1]. It contains at the moment about 70 plant species. The database provides an integrative, interactive and information-rich web graphical interface to display multidimensional data, and facilitate research and the discovery of functional NATs. Available information for the transcription factors (TF) for each species was retrieved from the Plant Transcription Factor Database. We have compared gene structure for wheat and related plant genomes.

Conclusion: The phenomenon of antisense transcription and miRNA interference need further annotation in new sequenced genomes. GO annotation and high-throughput small RNA sequencing data currently available will be integrated to investigate the biological function of such transcripts.

Acknowledgements: The work is supported in part by RFBR 11-04-01888, 12-04-00897; Integration projects SB RAS 21, 39, 130, Presidium RAS Programs 6.8, 30.29, No. 28.

References:

1. Chen et al. (2012) PlantNATsDB: a comprehensive database of plant natural antisense transcripts, *Nucleic Acids Res*, **40**(Database issue): D1187-93.

3D CHROMOSOME CONTACTS AND CHROMATIN INTERACTIONS REVEALED BY SEQUENCING

Orlov Y.L.*[1, 2, 3], Li G.[3], Auerbach R.[4], Sandhu K.S.[3], Ruan X.[3], Fullwood M.J.[3], Podkolodnyy N.L.[2], Afonnikov D.A.[1, 2], Liu E.[3], Wei C.L.[3], Snyder M.[4], 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;*
[4] *Stanford University, Stanford, CA, USA*
e-mail: orlov@bionet.nsc.ru
** Corresponding author*

Key words: *sequencing technologies, chromosome contacts, ChIP, PolII, 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) and similar techniques along with traditional in situ techniques have demonstrated that chromatin interactions can regulate transcriptional and epigenetic states.

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 including intra-genic, extra-genic and inter-genic 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].

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. Comparative analyses of different cell lines (such as MCF-7) imply that cell-line specific chromatin interactions could provide structural framework for cell-line specific transcription. 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.

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

Acknowledgements: The work is supported in part by A-STAR Singapore, RFBR 11-04-01888, 12-04-92702-IND, Integration projects SB RAS 21, 39, 130, Presidium RAS Programs 6.8, 30.29, No.28.

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.

SUPERCOMPUTER APPLICATIONS IN BIOINFORMATICS: SHARED FACILITY CENTER "BIOINFORMATICS" OF SIBERIAN BRANCH OF THE RUSSIAN ACADEMY OF SCIENCES

Orlov Y.L.*[1], Martyschenko M.K.[1], Afonnikov D.A.[1], Rasskazov D.A.[1], Fomin E.S.[1], Kuchin N.V.[2], Glinsky B.M.[2], Podkolodnyy N.L.[1, 2], Kolchanov N.A.[1]

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

[2] Institute of Computational Mathematics and Mathematical Geophysics, SB RAS, Novosibirsk, Russia

e-mail: orlov@bionet.nsc.ru

** Corresponding author*

Key words: *supercomputing, computer cluster, thigh-throughput computation, bioinformatics*

Motivation and Aim: Supercomputing is key component for modern bioinformatics researches. Activities covered under bioinformatics include comparative genomic analysis (as a result of the genomes sequencing projects), proteomic data analysis, phylogenetics and molecular evolution studies, protein structure modeling (including protein-ligand interactions), systems biology modeling, and data mining. All these areas require significant integrated computational resources for which GRID computing approaches are well suited. Integration of data sources coming from experimental installation (such as microscope and spectrometry devices) and data storage demands for such academic hubs as Novosibirsk Science Center require common supercomputer technological platform.

Methods and Algorithms: We have established Shared Facility Center "Bioinformatics" in the Siberian Branch of the Russian Academy of Sciences. The center uses HP-based supercomputer cluster of the Siberian Supercomputer Center (SSCC) to share computational resources for bioinformatics community.

Results: We have organized work on science sections on (1) computer genomics and transcriptomics, (2) computer proteomics, (3) modeling of molecular biological processes, (4) evolutionary bioinformatics (5) molecular dynamics, (5) mathematical problems of bioinformatics, (6) biological text mining, and (7) technical support of supercomputer researches. The center has web-portal for information exchange and work by science sections. For year 2011 we have 50% of computer load in Supercomputer Center SB RAS from bioinformatics group. Most time-consuming jobs are from computer genomics (de novo assembly of next generation sequencing data, up to 30% of total time), evolution research (simulation of evolution events in protein families) and molecular dynamics.

Conclusion: The Shared Facility Center "Bioinformatics" supports research biological problems demanding high-throughput computation. We report successful development of computer programs for biotechnology and genomics.

Availability: Shared Facility Center "Bioinformatics": http://biontomatics.bionet.nsc.ru

Siberian Supercomputer Center: http://www2.sscc.ru

Acknowledgements: support by the Russian Ministry of Education and Science (contracts No. 07.514.11.4011, 07.514.11.4023, 07.514.11.4003), RFBR 11-04-01888, Integration projects of SB RAS (No. 21, 36, 130), Integration Program of Presidium RAS 6.8.

TRANSCRIPTION FACTOR BINDING AND CHROMATIN MODIFICATIONS ANALYSIS BY CHIP SEQUENCING DATA

Orlov Y.L.*[1, 2], Li G. [3], Afonnikov D.A.[1, 2], Lim B.[3], Clarke N.[3], Huss M.[3], Gunbin K.V.[1], Ruan Y.[3], Podkolodnyy N.L.[1, 4], Chen M.[5], Ng H.-H.[3]

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

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

[3] *Genome Institute of Singapore, Singapore;*

[4] *Institute of Computational Mathematics and Mathematical Geophysics, SB RAS, Russia;*

[5] *Zhejiang University, Hangzhou, China*

e-mail: orlov@bionet.nsc.ru

** Corresponding author*

Key words: *transcription factors, ChIP-seq, sequencing, stem cells, computer genomics*

Motivation and Aim: High throughput sequencing technologies have enabled the identification of transcription factor (TF) binding sites in whole genomes. Important application is analysis of binding profiles in embryonic stem cells (ESCs). Somatic cells can be reprogrammed back to a pluripotent state by the combined introduction of transcription factors such as Oct4, Sox2, Klf4 and c-Myc. ChIP-seq data published recently and in the frames of ENCODE project allow construction of detailed genome binding maps. Such genome wide TF binding maps in mouse stem cells include Oct4, Sox2, Nanog, Tbx3, Smad2 as well as group of other factors. The signaling requirements for maintenance of human and murine embryonic stem cells (ESCs) differ considerably. Amongst the defined reprogramming factors, Oct4 is critical in inducing pluripotency. Integration of genome wide binding profiles for TFs allow find regulatory regions important for stem as no other factor has hitherto been shown to be able to substitute for Oct4, whilst Sox2, Klf4 and c-Myc can be replaced by other factors.

Methods and Algorithms: To meet issues of statistical analysis of genome ChIP-sequencing maps we developed computer program to filter out noise ChIP signals and find associations between TF binding affinity and number of sequence reads. We had search for homological regulatory regions in genomes of model organisms.

Results: Such TFs as Nr5a2, Eset, Smad2, PRDM14 uncover novel regulatory mechanisms for reprogramming. DNA accessibility (nucleosome depletion) facilitates TF binding as maybe measured by different techniques. Epigenetic modifications play important role in regulation of gene expression adding additional complexity to transcription network functioning. We have studied associations between different histone modifications using data for activating H3 histone marks H3K4me3, H3K4me1, H3K9ac and repressive histone marks H3K27me3, H3K9me3 together with RNA Pol II sites.

Conclusion: We found strong associations between activation marks and TF binding sites and present it qualitatively. Better prediction of TF binding could be achieved using chromatin modification data. Our data provide new insights into the function of chromatin organization in genome using integrative bioinformatics approaches.

Acknowledgements: The work is supported in part by Singapore A-STAR, RFBR 11-04-01888, 11-04-92712-IND, 12-04-92702-IND, Integration projects SB RAS 21, 39, 130, Presidium RAS Programs 6.8, 30.29, No.28.

ELICITING THE ROLE OF DIOXIN IN REGULATION OF THE GENES INVOLVED IN CYTOKINES SYNTHESIS BY MACROPHAGES

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

Key words: *macrophage, dioxin, TCDD, DRE, cytokines, promoters, transcription*

Motivation and Aim: Dioxin (and its prototype TCDD) is a one of the most potent toxicants. TCDD 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 biological, chemical, and physical stress, and therefore it is important to study the action of TCDD in these types of cells in order to decipher the possible mechanisms of TCDD immunotoxicity. On the cell level, the dioxin mediates gene expression via AhR/ARNT transcription complex activation, which binds to dioxin responsive elements (DRE) in the regulatory regions of the inducible genes. Previously, the TCDD–mediated modulation of the expression of genes encoding cytokines was shown experimentally [2]. For better understanding whether AhR/ARNT action is direct, or indirect through the immanent transcription factors, and to complete the list of cytokines synthesized in response to TCDD exposure we have performed the search of the putative DREs in the regulatory regions of the genes, encoding cytokines, expressed in activated macrophages.

Methods and Algorithms: The analysis of the gene network of macrophage activation by LPS and IFN-gamma was performed; the list of genes, involved in cytokine synthesis was formed; their regulatory regions were searched for DREs using the SITECON software package [3]. The experimental verification was performed using U937 human macrophages cell line and 2nM TCDD concentration.

Results: DRE sites have been detected and confirmed by EMSA in number of macrophageal cytokine gene promoters, with *IL12a, IL12b* and *IL-4* among others. Rt-PCR and ELISA experiments confirmed TCDD–mediated modulation of the expression of these genes.

Conclusion: TCDD can directly mediate the gene expression of the *IL12a, IL12b* and *IL-4* genes containing DREs in their regulatory regions, thus affecting the immune response.

Acknowledgements: This work was supported by RFBR (No. 12-04-01736).

References:

1. P.K. Mandal. (2005) Dioxin: a review of its environmental effects and its aryl hydrocarbon receptor biology. J Comp Physiol B. 175: 221-230.
2. C.F. Vogel et al. (2007) Modulation of the chemokines KC and MCP-1 by 2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD) in mice. *Arch Biochem Biophys.* 461:169-175.
3. D.Yu. Oshchepkov et al. (2004) SITECON: a tool for detecting conservative conformational and physicochemical properties in transcription factor binding site alignments and for site recognition. *Nucl. Acids Res.* **32:** W208–212.

PROMOTERS OF THE GENES ENCODING THE KEY TRANSCRIPTION FACTORS IN THE INFLAMMATORY RESPONSE CONTAIN BINDING SITES FOR ARYL HYDROCARBON RECEPTOR

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

Key words: *macrophage, dioxin, TCDD, transcription factor, promoters, binding sites*

Motivation and Aim: 2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD) is the most toxic among the dioxin xenobiotics, able to induce a broad spectrum of biological responses, including also cancer and immunotoxicity [1]. TCDD mediates gene expression via AhrR/Arnt transcription complex activation, which binds to dioxin responsive elements (DRE) in the gene regulatory regions. Inflammatory conditions in some organs increase the risk of cancer, on one hand [2]. On the other hand, several studies have shown that TCDD acts as a stimulator of inflammatory cytokines [3], and our subsequent analysis [4] showed the possibility of indirect regulation of these cytokines via intrinsic transcription factors, containing DREs in its regulatory regions. The list of the genes of these TFs contains also Stat3 and subunits of NFκB, a key transcription factors in the inflammatory response.

Methods and Algorithms: The regulatory regions of the genes, encoding key TFs and it's subunits were searched for DREs using the SITECON software package, designed for searching for the transcription factors binding sites in genomic sequences [5]. The experimental verification was performed using U937 human macrophages cell line and 2nM TCDD concentration. EMSA and Rt-PCR experiments were performed to test sites functionality.

Results: DRE sites have been detected and confirmed by EMSA in number of transcription factor gene promoters, with subunits of NFκB Rel and RelA among others. Rt-PCR experiments confirmed TCDD–mediated modulation of the expression of these genes.

Conclusion: TCDD–mediated modulation of the genes NFκB subunits Rel and RelA could be a possible pathway underlying dioxin-induced cancer risk.

Acknowledgements: This work was supported by RFBR (No. 12-04-01736).

References:

1. P.K. Mandal. (2005) Dioxin: a review of its environmental effects and its aryl hydrocarbon receptor biology. *J Comp Physiol B*. 175: 221-230.
2. A. Mantovani (2010) Molecular pathways linking inflammation and cancer. *Curr Mol Med.* 10(4):369-73.
3. C.F. Vogel, *et al.* (2005) Induction of proinflammatory cytokines and C-reactive protein in human macrophage cell line U937 exposed to air pollution particulates. *Environ Health Perspect.* 113:1536-1541.
4. Furman DP, *et al.* (2009) Promoters of the genes encoding the transcription factors regulating the cytokine gene expression in macrophages contain putative binding sites for aryl hydrocarbon receptor. *Comput Biol Chem.* 33(6):465-8.
5. 5. D.Yu. Oshchepkov, *et al.* (2004) SITECON: a tool for detecting conservative conformational and physicochemical properties in transcription factor binding site alignments and for site recognition. *Nucl. Acids Res.* 32: W208–212.

DEPPDB – A PORTAL FOR ELECTROSTATIC AND OTHER PHYSICAL PROPERTIES OF NATURAL GENOMES

Osypov A.A.*, Krutinin G.G., Krutinina E.A., Kamzolova S.G.
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 potential, physical properties, data integration, genomics*

Motivation and Aim: Electrostatic and other physical properties of genome DNA influence its interactions with different proteins, in particular the regulation of transcription by RNA-polymerases. 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, which are 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 new program package [2,5]. Calculations of other physical properties are based on the di- and trinucleotide content. Different smoothing and cross-correlation algorithms are applied.

Results: Currently, the database contains all the completely sequenced bacterial, viral, mitochondrial and plastids genomes according to the NCBI RefSeq [3] as well as some model eucariotic genomes. Data for promoters, regulation sites, binding proteins *etc.* are incorporated from Dbs and literature. All the data are fully integrated and 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 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 complete genomes.

Availability: The database [5] 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. D. K. Pruitt, T. Tatusova, and D. R. Maglott. (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins, NAR, 35, D61–D65.
4. 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

ELECTROSTATIC AND OTHER PHYSICAL PROPERTIES OF NATURAL GENOMES AND THEIR ELEMENTS

Osypov A.A.*, Krutinin G.G., Krutinina E.A., Kamzolova S.G.
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 potential, physical properties, data integration, genomics*

Motivation and Aim: Many physical properties of genome regulator regions, rising from the same (basically) sequence composition, may complement each other in their functionality, providing possible refined regulatory effects due to smaller differences in their composition-dependent formation. Electrostatic profile does not correspond one-to-one to the sequence due to the massive contextual effects and the surroundings of a sequence longer than a typical conservative regulation region (*i.e.* >10 b.p.), can shift potential in its center by more than its own typical spread. The number of 10bp is 4^{10}=1048576. Largest known bacterial genomes are about 10 Mbp long and the sequences distribution in them is very non-uniform, (in fact, the *E.coli* 4.6 Mbp genome has gaps even in 7-bp oligomers), *many of the sequences are not present in the natural genomes for studying directly possible biological effects of their physical properties*, not to speak of any statistics. This may be considered as a "mathematical proof" of the necessity of DNA electrostatic potential and other physical properties studies *in silico* by the bioinformatics methods.

Methods and Algorithms: DEPPDB – DNA Electrostatic and other Physical Properties Database and its tools [1,2] were used to carry out the analysis.

Results: All the completely sequenced bacterial, viral, mitochondrial and plastids genomes according to the NCBI RefSeq, as well as some model eukariotic genomes were examined according to the averaged value of more than 125 physical and conformation properties. Different types of properties values distributions according to the genomes GC content are observed and classified. Distributions of different types of artificial DNA sequences are calculated, revealing some non-linear dependences of specially arranged motifs compared to common statistics-based natural distributions. Also observed and analyzed are data for transcription starting sites, promoters, regulation (binding proteins) sites, terminators *etc*. Several pronounced peculiarities are revealed and discussed.

Availability: The database [1,2] 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. 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

ORGANIZATION OF XENOBIOTIC-METABOLIZING SYSTEM PHASE 1 IN *OPISTHORCHIS FELINEUS* (TREMATODA, PLATYHELMINTHES)

Pakharukova M.Y.[1], Ershov N.I.[1], Vavilin V.A.[2], Vorontsova E.V.[1], Katokhin A.V.[1], Duzhak T.G.[3], Merkulova T.I.[1], Mordvinov V.A.[1]

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

[2] *Institute of Molecular Biology and Biophysics, SB RAMS, Novosibirsk, Russia;*

[3] *International Tomography Center, SB RAS, Novosibirsk, Russia*

Infection with the human liver fluke *Opisthorchis felineus* is a serious public health problem in Russia and other Eastern Europe countries. *O. felineus* infests human bile ducts of the liver and the pancreas, has significant pathogenicity, and causes complications including cholecystitis, liver abscesses and other diseases.

The aim of this work was to identify and to characterize phase I of xenobiotic-metabolizing system, cytochrome P450 (CYP450, CYP) mRNA in *O. felineus*.

To reconstruct *O. felineus* CYP450 mRNA sequence we used available Platyhelminth genome and transcriptome data. The only one cytochrome P450 has been identified in *O. felineus* and in other various species of parasitic flatworms.

We have cloned and sequenced CYP450 mRNA of *O. felineus*. The CYP amino acid sequence was reconstructed and confirmed by MS/MS. The protein sequence contains conserved functional domains, such as in other eukaryotic microsomal CYP450 enzymes participating in the biotransformation of xenobiotics and drugs.

In contrast to the low level of amino acid sequence homology to other eukaryotic CYPs (20-24%) 3D model of the *O. felineus* CYP450 demonstrates high conformational similarity with mammalian CYP2 subfamily structures. The observed conformational similarity between proteins of biotransformation of xenobiotics *O. felineus* and humans indicates a high probability of coincidence of their functions.

Indeed, we have demonstrated a monooxygenase CYP activity in fluke. The activity was similar to the mammalian CYP2E1 (HPLC) and mammalian CYP2B (MS). The level of *O. felineus* CYP mRNA expression (Real-time PCR) in maritae (adult stage in definitive mammal host) was constitutive and significantly higher than in other life stages. This indicates an important role of the biotransformation enzyme in biochemistry of the parasite and in the host-parasite relationships in mammalian host.

This work is supported in part by RFBR grants (N 11-04-01333a and N 11-04-90490ukr-f-a) and SB RAS Research Partnership grant N 19.

GENETALK: AN EXPERT EXCHANGE PLATFORM FOR ASSESSING RARE SEQUENCE VARIANTS IN PERSONAL GENOMES

Parkhomchuk D.V.[1], Kamphans T.[2], Heinrich V.[1], Krawitz P.*[1]
[1] *Department of Medical and Human Genetics, Charite, Universitetsmedizin Berlin, Augustenburger Platz 1, 13353 Berlin, Germany;*
[2] *GeneTalk, Finckensteinallee 84, 12205 Berlin, Germany*
e-mail: peter.krawitz@gene-talk.de
** Corresponding author*

Key words: *rare genetic diseases, variants filtering and annotation*

Motivation and Aim: Next-generation sequencing (NGS) has become a powerful tool in personalized medicine. Exomes or even whole genomes of patients suffering from rare diseases are screened for sequence variants. After filtering out common polymorphisms, the assessment and interpretation of detected personal variants in the clinical context is an often time consuming effort.

Methods and Algorithms: We have developed GeneTalk, a web-based platform that serves as an expert exchange network for the assessment of personal and potentially disease relevant sequence variants. GeneTalk assists a clinical geneticist who is searching for information about specific sequence variants and connects this user to other users with expertise for the same sequence variant.

Results: GeneTalk is an exchange platform that allows users to look for variant specific information and makes human expertise searchable. Any sequence variant with respect to the human reference genome, based on the GRChR37 assembly, is annotatable and the user decides to whom this annotation is visible. A user could link to scientific articles that are relevant in context with a certain variant or that even provide evidence that a mutation is disease causing. A second user might comment on this annotation to express his concern because he views the detected variant as an technical artefact e.g. due to pseudogenes. A third user might state that he has seen patients with this genotype and is not sure about the statistical significance of the association with the phenotype. All annotations and comments of GeneTalk users about a certain genomic position can be read like a locus specific conversation thread. The trustworthiness of an annotations can be rated by users as well as the likelihood of a mutation to be disease causing.

Conclusion: GeneTalk provides an intuitive web-based interface for geneticists that analyze human sequence variants. GeneTalk is a platform for efficient knowledge management of genetic variants and simplifies the scientific discussion and interpretation especially of rare mutations.

Availability: GeneTalk is available at www.gene-talk.de.

References:

1. Bamshad M.J., et al. (2011) Exome sequencing as a tool for Mendelian disease gene discovery. *Nat Genet Rev,* **12:** 745-55.
2. Robinson, P.N., et al. (2011) Strategies for exome and genome sequence data analysis in disease-gene discovery projects. *Clin Genet,* **80:** 127-32.

PUTracer: A NOVEL METHOD FOR IDENTIFICATION OF CONTINUOUS-DOMAINS IN MULTI-DOMAIN PROTEINS

Parsa M.[1], Pashandi Z.[2], Mobasseri R.[2], Arab S.S.*[2]
[1] *Institute for Research in Fundamental Sciences IPM, Tehran, Iran;*
[2] *Tarbiat Modares University TMU,Tehran, Iran*
e-mail: sh.arab@modares.ac.ir
** Corresponding author*

Key words: *proteins, continues-domains, accessible surface area, energy of hydrogen bonds, top-down approach, recursive*

Motivation and Aim: Computer-assisted assignment of protein domains is considered as an important issue in structural bioinformatics. Exponential increasing in the number of known three-dimensional structure of proteins and the significance role of proteins in biology, medicine and pharmacology still illustrate the necessity of a reliable method to detect automatically structural domains as protein units.

Methods and Algorithms: For this aim, we develop a program based on the accessible surface area (ASA) and the energy of hydrogen bonds in backbone residues (EHB). PUTracer builds on the features of a fast top-down approach to cut a chain or domain into the new domains with minimal change in ASA as well as EHB, and without destruction of disulfide bonds at every recursive step.

Results: Performance of program was assessed by a comprehensive benchmark dataset of 124 protein chains which is based on agreement among experts (e.g., CATH, SCOP) and was expanded to include structures with every type of domain combinations. Equal number of domain and at least 90 % agreement in critical boundary accuracy were considered as correct assignment conditions. PUTracer assigned domains correctly in more than 82% of proteins.

Conclusion: Although, low critical boundary accuracy in 18 % of proteins leads to the incorrect assignments, adjusting the scales is likely to improve the performance up to 92 %. We discuss here the successes or failure of adjusting the scales with provided evidences.

Availability: PUTracer is available at http://bioinf.modares.ac.ir/software/PUTracer/

A DISCRETE DYNAMICAL SYSTEM ON A DOUBLE CIRCULANT WITH AN ADDITIVE FUNCTION OF THE VERTICES

Perezhogin A.L., Imangaliyeva Zh.G.*
Novosibirsk State University, Novosibirsk, Russia
e-mail: zhamiga@yahoo.com
** Corresponding author*

Key words: *gene network, discrete dynamical system, carrier graph, functional graph, working points, fixed points, cycles*

Motivation and Aim: Models of regulatory loops with positive and negative feedbacks are being actively studied. These two types of loops make it possible to maintain a certain functional state or slide to another state of a gene network [1]. We consider one of the methods for description and modeling of gene networks - in terms of the discrete models of two connected regulatory loops functioning.

Methods and Algorithms: A gene network is presented as a connected digraph with n vertices, identified by the proteins, and a set of arcs, associated with the regulatory relations. Each vertex is attributed by a variable taking on a value of either 0 or 1 that defines the concentration of protein in the appropriate vertex. A carrier graph consisting of two double circulant subgraphs $G_{n,2}$ [2] is considered and the functionality of the graph is defined by the additive threshold function. In the study of the functioning of such dynamic system, we use the methods of discrete analysis.

Results: The approach to the complexity analysis for the discrete models of gene networks on the basis information on their functioning is proposed in the work and the theoretical and computer analysis for various parametric data of the regulatory loops is presented. All the vertices with loops (that correspond to fixed conditions of the initial network), dangling vertices, vertices included in a component of connectivity with zero vertex are described. The problem of describing all the vertices of the functioning of non-suspended was reduced to the more simple dynamical system.

Conclusion: Investigation of the regulatory loops functioning and their influence on each other represented here provides an opportunity to understand the regulatory mechanisms of the processes under the control of gene networks and possibility for a directional impact on them.

References:

1. V.A. Likhoshvai, V.P. Golubyatnikov, G.V. Demidenko, A.A. Evdokimov, S.I. Fadeev. (2008) Theory of the gene networks (Russian). In: *System computerized biology*, Eds. N.A. Kolchanov and S.S. Goncharov, 430–576, Novosibirsk, SB RAS.
2. E.D. Grigorenko, A.A. Evdokimov, V.A. Likhoshvai, I.A. Lobareva. (2005) Fixed points and cycles of the automata mapping for the gene network simulation (Russian). *Herald of Tomsk State University,* **14**: 206 - 212.

THE DISCRETE DYNAMIC SYSTEM ON A DOUBLE CIRCULANT WITH DIFFERENT FUNCTIONS AT THE VERTICES

Perezhogin A.L., Nazhmidenova A.M.*
Sobolev Institute of Mathematics, SB RAS, Novosibirsk, Russia
e-mail: deviliona@yandex.ru
** Corresponding author*

Key words: *gene network, discrete model, regulatory loop, functional graph, cycles, fixed point and pendant vertices*

Motivation and Aim: The regulatory loops with positive and negative feedbacks play an important role in gene networks. These two types of loops make it possible to maintain a certain functional state or slide to another state of a gene network [1]. We consider one of the methods for description and modeling of gene networks - in terms of the discrete models of two connected regulatory loops with different feedbacks.

Methods and Algorithms: The regulatory loop is a connected digraph with *2n* vertices composing of two circulants $G_{n,k}$, which corresponding vertices are conjugate. Circulant digraphs $G_{n,k}$, where (k–1) is a number of inputs, are considered [2]. Variables are taking values 0 or 1. The functioning is set by multiplicative mapping on the first circulant and additive – on another. The existence of fixed points of mappings means the possibility of gene networks to be in stationary states. Pendant vertices of the functional graph characterize inaccessible states. Cycles mean the ability of gene network to return to the initial state.

Results: We propose an approach to the complexity analysis for the discrete models of gene networks on the basis information on their functioning. In the case $k = 2$ the theorems characterizing structural properties, fixed points, pendant vertices and cycles of length two of the functional graphs were received. In particular, the explicit formulas for the number of fixed points and pendant vertices were found. Also the recurrent relation for the number of fixed points for any k was obtained, and the asymptotic behavior of this number was described.

Conclusion: Investigation of the regulatory loops functioning and their influence on each other represented here provides an opportunity to understand the regulatory mechanisms of the processes under the control of gene networks and possibility for a directional impact on them.

References:

1. V.A. Likhoshvai, V.P. Golubyatnikov, G.V. Demidenko, A.A. Evdokimov, S.I. Fadeev. (2008) Theory of the gene networks (Russian). In: *System computerized biology*, Eds. N.A. Kolchanov and S.S. Goncharov, 430–576, Novosibirsk, SB RAS.
2. E.D. Grigorenko, A.A. Evdokimov, V.A. Likhoshvai, I.A. Lobareva. (2005) Fixed points and cycles of the automata mapping for the gene network simulation (Russian). *Herald of Tomsk State University,* **14**: 206 - 212.

RECOMBINATION OF MOBILE GENETIC ELEMENTS AS POSSIBLE SOURCE OF NEW ISSR-PCR MARKERS

Pheophilov A.V.*, Glazko V.I.
Russian State Agrarian University – Moscow Agricultural Academy named after K.A. Timiryazev, Moscow, Russia
e-mail: foton87@yahoo.com
** Corresponding author*

Key words: *inverted repeats, retrovirus, ISSR-PCR, mobile genetic elements*

Motivation and Aim: Due to success in sequencing and annotating genomes of agricultural animals, genomic scanning based on the multiloci markers becomes more and more important, thus making finding out properties of polymorphism of various genomic elements necessary for their effective applications in controls of gene pool dynamics of farm animals. We have carried out comparative investigations of spectra of genomic DNA fragments acquired by means of ISSR-PCR from different cattle, horse and sheep breeds and found features specific for species as well as for breeds. To reveal possible mechanisms of such process we sequenced one DNA fragment, the presence of which was species specific for amplification products of spectra, obtained in PCR of horse genomic DNA using $(AG)_9C$ as primer.

Methods and Algorithms: The sequencing of a DNA fragment of 416 bp length flanked by an inverted repeat $(AG)_9C$ which was observed only in horses was carried out. The DNA sequencing was carried out in the Inter-Institute Center of collective using "GENOME" of IMB Russian Academy of Sciences {{http://www.genome-centre.narod.ru/}} organized with the support of RFBR. The search for homology with female thoroughbred horse stored in GenBank by using BLASTn algorithms didn't show homology of the whole fragment. Thus making us use RepeatMasker and CENSOR tools {{http://www.repeatmasker.org/}}, {{http://www.girinst.org/censor/}} to search it in repeats database.

Results: Three fragments of homology to different mobile elements were revealed in the query. From 1st to 39th bp – to nonautonomous DNA transposon from *Danio rerio;* from 46th to 166th – to part of endogenous retrovirus LTR ERV 3, firstly described in human genome; from 182nd to 416th – to part of LTR ERV1, putative non-autonomous ERV1-type endogenous retrovirus, revealed only in horse genome [1]. It turned out that the latter part of 237 bp length is very frequently localized in horse genome and its frequency of localization in different horse chromosomes statistically reliable (r=0.93) correlates with the length of each chromosome including X chromosome.

Conclusion: The data acquired evidence that some ISSR-PCR markers taking part in interspecie and perhaps intraspecie differentiation may originate from recombinations between evolutionary "old" and relatively "younger" mobile genetic elements.

References:

1. J. Jurka. (2008) Putative non-autonomous ERV1-type endogenous retrovirus from horse, *Repbase Reports*, Vol.8, No.5: 601.

COMPARATIVE GENOME ANNOTATION OF TRYPANOSOMATIDS

Pintus S.S.*[1], Serrano M.G.[1], Alves J.M.[1, 2], Matveyev A.[1], Sheth N.[1], Lara A.[1], Lee V. [1], Koparde V.N.[1], Rivera M.C.[1], Voegtly L.J.[1], Arodz T.J.[1], Maia da Silva F. [2], Camargo E.P. [2], Teixeira M.M.G.[2], Buck G.A.[1]

[1] Virginia Commonwealth University, Center for the Study of Biological Complexity and the Departmnent of Microbiology and Immunology, Richmond VA, USA;

[2] Departamento de Parasitologia, Universidade de São Paulo, São Paulo, SP, Brazil

email: sspintus@vcu.edu

** Corresponding author*

Key words: *Comparative genomics, trypanosomatids, rare tropical diseases, high throughput sequencing, genome annotation, Assembling the Tree of Life project*

Motivation and Aim: Trypanosomatids are a group of exclusively parasitic kinetoplastid protozoa, infecting primarily insects. Life cycles of some trypanosomatid species, e.g. *T. brucei*, *T. cruzi* and various species of *Leishmania* involve a secondary human host, resulting in major deseases in humans, such as sleeping sickness, Chagas disease and leishmaniasis. Here we present an approach and results for high throughput sequencing and comparative annotation of several genomes of trypanosomatids, including the genera *Trypanosoma*, *Leishmania* and *Crithidia*.

Methods and Algorithms: The whole genome sequencing was done using Roche 454 pyrosequencing and Illumina (Solexa) technology. The assemblies of the genomes were performed *de novo* and improved using various software, such as Newbler (Roche), CLC Assembly Cell (CLC Bio), Velvet (EMBL-EBI), Minimus2 (AMOS consortium), etc.

In addition to recently developed ASGARD software [1], we have developed an informatics pipeline to annotate genes from the genomes under study. The pipeline includes such steps as prediction of protein coding genes, annotation of the gene orthologs and protein families, tRNA genes prediction, etc.

To assess the completeness of the assemblies in terms of genes annotation, we estimated the presence of trypanosomatid orthologs in genomes under study. The sets of trypanosomatid orthologs were recently created applying OrthoMCL software (PCBI, University of Pennsylvania) on the annotated *T. cruzi* CL Brenner, and *L. major* genomes. The ortholog presence tests were performed using the recently developed tool which implements Blast searches against the ortholog sets.

Calling orthologs from different incomplete subsets of reads allows draft estimating of the completeness of the genome assembly, while comparison between ortholog calls from annotated genes and contigs, allows identification of certain misassembled genes, as well as giving insights for trypanosomatid genome evolution.

Acknowledgments: The work was supported by grant Assembling the Tree of Life: Phylum Euglenozoa NSF DEB 0830056 Buck (PI) 09/15/08 – 09/14/13

References:

1. J.M. Alves, G.A. Buck. (2007) Automated system for gene annotation and metabolic pathway reconstruction using general sequence databases, Chemistry & Biodiversity, 4:2593-2602.

COMPUTATIONAL EVALUATION OF IMPACT OF AMINO ACID SUBSTITUTION p.W172C ON STRUCTURE AND FUNCTION OF GAP-JUNCTION PROTEIN CONNEXIN 26 AND ITS ASSOCIATION WITH HEARING IMPAIRMENT

Pintus S.S.[1], Bady-Khoo M.S.[1, 2], Posukh O.L.*[1, 3]
[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia; [2] Republican Hospital #3, Kyzyl, Russia; [3] Novosibirsk State University, Novosibirsk, Russia
** Corresponding author: posukh@bionet.nsc.ru*

Motivation and Aim: Mutations in the *GJB2* gene, encoding the gap-junction protein connexin 26 (Cx26), are the most common cause of non-syndromic deafness. Transmembrane protein Cx26 forms intercellular channels that permit the exchange of ions and small molecules between adjacent cells. Defects of Cx26 lead to the disturbance in ion homeostasis of inner ear endolymph which results in hearing impairment. To date, about 200 different pathogenic mutations, polymorphisms and changes with unknown relation to the disease in *GJB2* gene have been reported («Connexins and Deafness Homepage»: http://davinci.crg.es/deafness/). We analysed nucleotide sequences of the *GJB2* gene entire coding region in 90 deaf patients of Tuvinian ethnicity (the Tuva Republic, Russia) and revealed sequence alterations: c.516G>C (p.W172C), c.235delC (p.L79fs), c.109G>A (p.V37I), c.299_300delAT (p.H100fs), c.79G>A (p.V27I), c.341A>G (p.E114G), and c.571T>C (p.F192L). Most of them are known pathogenic mutations or polymorphisms. Interestingly, we observed high prevalence of a nonsynonymous substitution c.516G>C (p.W172C) (79.4% out of all mutant chromosomes) which previously rarely detected in deaf patients (Posukh et al., 2005; Tekin et al., 2010). The aim of this study is *in silico* evaluation of the p.W172C impact on structure and function of protein Cx26 and its association with hearing impairment.

Methods and Algorithms: The PolyPhen (http://genetics.bwh.harvard.edu/pph/) and PolyPhen2 (http://genetics.bwh.harvard.edu/pph2/) are the tools which predict possible impact of an amino acid substitution on the structure and function of a human protein using straightforward physical and comparative considerations. Both programs were used for prediction of the p.W172C effect on structure and function of human Cx26 (http://www.uniprot.org: CXB2_HUMAN). Multiple sequence alignment of the Cx26 amino acid sequences of different species was performed by ClustalX.

Results: The p.W172C change is located in the second extracellular domain of Cx26. Its effect on the protein Cx26 was assessed using the PolyPhen where a PSIC (position-specific independent counts) score difference between W and C was found to be 2.669 that classified W172C as probably damaging mutation. According to the PolyPhen2 prediction model HumVar the W172C was also classified as probably damaging mutation (p=0.901). Multiple sequence alignment of the Cx26 amino acid sequences in different species (*H. sapiens, O. anatinus, M. domestica, H. glaber, C. porcellus, M. gallopavo*) by ClustalX revealed high conservative W at position 172 in connexin 26.

Conclusion: Based on evolutionarily conservation of W at position 172 in connexin 26 in many species and the PolyPhen analysis, p.W172C (c.516G>C) was considered to be probably pathogenic alteration and associated with hearing impairment. Substitution of an aromatic non-polar tryptophan at position 172 on small polar cysteine probably results in impairment of connexin protomers connection or opening-closing mechanism of intercellular Cx-channels.

Acknowledgements: The study was supported by RFBR grant #11-04-01221-a.

HIGH PERFORMANCE COMPUTING IN BIOINFOMRANTICS: CASE STUDIES

Podkolodnyy N.L.*[1, 2], Demenkov P.S.[1], Gunbin K.V.[1], Orlov Y.L.[1], Fomin E.S.[1], Alemasov N.A.[1], Kazantsev F.V.[1], Vishnevsky O.V.[1], Ivanisenko V.A.[1], Afonnikov D.A.[1], Kuchin N.V. [2], Glinsky B.M.[2], Kolchanov N.A. [1]
[1] *Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia;*
[2] *Institute of Computational Mathematics and Mathematical Geophysics, SB RAS, Novosibirsk, Russia*
e-mail: pnl@bionet.nsc.ru
** Corresponding author*

Key words: *parallel computing, thigh-throughput computation, bioinformatics*

Motivation and Aim: The large-scale researches in bioinformatics and computer systems biology requires the use of high performance computing due to the explosive growth in the volume and diversity of molecular biological data obtained during the execution of projects for the sequencing of genomes, as well as the need to calculate complex models of biological objects of macromolecules and drugs, gene networks. This determines the significance of developing parallel versions of the most computationally expensive tasks of bioinformatics focused on the use of either large computational cluster architecture with MPI, or for some tasks, the use of hardware accelerators.

Methods and Algorithms: There are several approaches for acceleration of computational experiments: usage of high-throughput computation clusters or supercomputers with shared memory, parallelization by data, parallelization by tasks and software packages, parallelization at instruction level, usage of special processors (apparatus acceleration) GPU (Graphics Processing Unit), FPGA (Field Programmable Gate Array), usage of hybrid architecture joining CPU together with special processors GPU or FPGA. We have work in the frames of shared facility center "Bioinformatics" technically based at the Siberian Supercomputer Center (SSCC). The HPC cluster with total peak performance 115 TFlops has hybrid architecture and consists of cluster of NCC-30T (platform BL2h220c hp) with 576 Intel Xeon processors E5450/E5540/X5670 (2688 cores - peak performance 30 teraflops) and hybrid cluster based on 40 servers HP SL390s G7 (80 CPU X5670 - 480 cores) each with three graphical accelerators (GPU) NVidia Tesla M2090 (61 440 cores).

Results and conclusion: The parallel version of programs for solving various classes of tasks of computer genomics, proteomics, molecular dynamics and mathematical modeling of molecular-genetic processes has been developed. This paper presents a series of numerical experiments performed using HPC, in particular, assembly of short reads in metagenomic studies, analysis of protein structure under extreme conditions of high pressures and temperatures using molecular dynamics study, analysis of changes in the thermal stability of proteins under different mutations, using the estimates of free energy, obtained by molecular dynamics, the discovery new functional regulatory motifs in genomic sequences, promoter prediction, ChIP-on-chip data analisys and others.

Acknowledgements: support by the Russian Ministry of Education and Science (contracts No. 07.514.11.4011, P857), Integration projects of SB RAS (No. 136, 21, 39), Program "Genomics. Proteomics. Bioinformatics" SB RAS.

DISTRIBUTED RESTFUL-WEB-SERVICES FOR THE RECONSTRUCTION AND ANALYSIS OF GENE NETWORKS

Podkolodnyy N.L.*[1, 2], Semenychev A.V.[1], Borovsky V.G.[1], Rasskazov D.A.[1], Ananko E.A.[1], Ignatieva E.V. [1], Podkolodnaya N.N.[1], Podkolodnaya O.A.[1]
[1] *Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia;*
[2] *Institute of Computational Mathematics and Mathematical Geophysics, SB RAS, Novosibirsk, Russia*
e-mail: pnl@bionet.nsc.ru
** Corresponding author*

Key words: *RESTful-Web Services, gene networks, graph analysis*

Motivation and Aim: Graph-based biological networks models represent the structure of the functional relationships between molecular entities such as gene, protein and small compounds and provide a suitable framework for integrating and analyzing omics-data. This paper focuses on the data integration and reconstruction of the different variants of the structural representation of molecular genetic systems under study as a first step in creating a dynamic model of molecular genetic systems. This is the most important and poorly formalized stage of gene network modeling which involves building a formal description of the system on the basis of current knowledge integration and understanding.

Methods and Algorithms: The REST style web services became a popular alternative of SOAP based services and they are considered lighter and easier to use. Using the RESTful web services technology provide the possibility to develop the distributed system for integration of large, heterogeneous biological datasets from multiple sources with the such important property as scalability, statelessness, cache ability, uniformity.

Results and conclusion: The distributed system for gene network reconstruction and analysis of gene network structure based on RESTful web services technology was developed. The gene network reconstruction is based on integration of information on molecular entities, molecular interactions, gene regulation and molecular pathways accumulated in databases GeneNet, TRRD, KEGG, SWISS-PROT, Pathway, IntAct, InterPro, MINT, bioCyc, BRENDA and other information sources, in particular, web services to ANDCell computer system which includes a database of knowledge and facts extracted automatically from PubMed and web services to Protein Structure Discovery system for predict different kind of molecular interactions. The most important methods for analyzing the structure of gene networks are realized: calculation of the integral characteristics of genetic networks based on graph theory, search for elementary structural motifs that are the building blocks of gene networks and analyze their distribution in different types of gene networks, search and analysis of the cycles (negative or positive feedbacks) in gene regulatory networks, analysis of strongly connected components and their condensation in gene networks, search of gene network clusters and reduction of the metabolic pathways or gene networks as a general approach to analysis and structured modeling of complex molecular-genetic systems, and other.

Acknowledgements: This work was supported by Ministry of Education and Science of the Russian Federation State (Contract № 07.514.11.4023).

MULTISTATE ORGANIZATION OF TRANSMEMBRANE HELICAL PROTEIN DIMERS GOVERNED BY THE HOST MEMBRANE

Polyansky A.A.*[1, 2], Volynsky P.E.[1], Efremov R.G.[1], Shemyakin M.M.[1], Ovchinnikov Yu.A.

Institute of Bioorganic Chemistry, Russian Academy of Sciences, Moscow, Russia;
[2] Department of Structural and Computational Biology, Max F. Perutz Laboratories, University of Vienna, Campus Vienna Biocenter 5, Vienna, AT-1030, Austria
** Corresponding author*

Keywords: *prediction of protein structure, molecular dynamics simulations, free energy calculations, molecular hydrophobicity potential, protein-membrane interactions, receptor tyrosine kinases, epidermal growth factor receptors*

Motivation and Aim: Association of transmembrane (TM) helices taking place in the cell membrane has an important contribution into the biological function of bitopic proteins, among which receptor tyrosine kinases represent a typical example and very potent target for medical applications. Since this process is driven by many factors (i.e. primary structures of TM domains and juxtamembrane regions, composition and phase of the local membrane environment, etc.), it is still far from being fully understood.

Methods and Algorithms: We have used original modeling approach (so-called PREDDIMER [1]), which allows prediction of TM helical oligomers from their primary sequences based on quantitative estimations of complementarity between geometrical and polar properties of helical surfaces. We have estimated the free energy of association of several predicted dimers in in full-atom explicit lipid bilayers composed of phosphocholine lipids with different acyl chains, using umbrella sampling techniques with the mean force integration.

Results: We present a computational modeling framework, which we have applied to systematic consideration of dimerization for 18 TM helical homo- and heterodimers of different bitopic proteins, including the family of epidermal growth factor receptors. For this purpose, we have developed a novel surface-based modeling approach, which is able not only to predict some particular conformations of TM dimers displaying good agreement with the experiment, but also provides screening of their conformational heterogeneity together with simple estimation of the dimerization efficiency. To elucidate a putative role of the environment in a selection of a particular conformation, we have employed full-atomic MD simulations of several of the predicted dimers in different model membranes. Analysis of about 20 μs of MD statistics clearly shows that each particular bilayer preferentially stabilizes one of possible conformations of a dimer, and that the energy gain depends on interplay between structural properties of the protein and the membrane.

Conclusions: Our results suggests a multistate organization of TM helical dimers in heterogeneous membranes, and emphasizes an importance of consideration of their conformational variability to design potent selective modulators of dimerization acting on pharmaceutically relevant targets in the natural medium of cell membranes.

Refereces:

1. A. A. Polyansky, P. E. Volynsky, R. G. Efremov. (2011) *Adv. Protein Chem. Struct. Biol.*, **83**, 129-161.

EPIGENETIC STATUS AND QUANTITATIVE CHARACTERISTICS OF CIRCULATING DNA IN LUNG CANCER

Ponomaryova A.A.*[1], Rykova E.Y.[2], Cherdyntseva N.V.[1], Skvortsova T.E.[2], Dobrodeev A.Y.[1], Zav'yalov A.A.[1], Tuzikov S.A.[1], Vlassov V.V. [2], Laktionov P.P. [2]
[1] *Cancer Research Institute, SB RAMS, Tomsk, Russia;*
[2] *Institute of Chemical Biology and Fundamental Medicine, SB RAS, Novosibirsk, Russia*
e-mail: anastasia-ponomaryova@rambler.ru
** Corresponding author*

Key words: *circulating DNA, methylation, tumor suppressor genes, lung cancer*

Motivation and Aim: Analysis of concentration and tumor suppressor genes methylation of cirDNA in lung malignancies and chronic obstructive pulmonary disease.

Methods and Algorithms: Blood samples were taken from 32 healthy subjects (HS), 22 patients with chronic obstructive pulmonary disease (COPD) and 55 untreated patients with non-small cell lung cancer (NSCLC) before treatment. The cell-surface-bound cirDNA (csb-cirDNA) fractions were obtained by successive treatment with PBS/EDTA and trypsin solutions. The copy number of ACTB gene and LINE-1 repetitive element was measured by quantitative real-time PCR. Concentration of methylated and unmethylated RASSF1A and RARB2 tumor suppressor genes circulating in blood was quantified by methylation-specific PCR and methylation index (IM) was calculated as IM = 100 × [copy number of methylated /(copy number of methylated + unmethylated gene)] for cirDNA and csb-cirDNA.

Results: Using PCR assays the significantly decreased concentrations of csb-cirDNA were shown in the blood of NSCLC and COPD patients compared with HS ($P<0.05$ – for ACTB and $P<0.01$ – for LINE-1, Mann-Whitney U test). IM values for RASSF1A and RARB2 genes were significantly elevated in cirDNA from NSCLC patients compared with HS (39% vs 19% in csb-cirDNA, 45% vs 21% in plasma cirDNA for RASSF1A; 35% vs 11% in csb-cirDNA, 51% vs 17% in plasma cirDNA for RARB2; $P<0.05$). If at least one from RASSF1A or RARB2 IM exceeded the cut-off values NSCLC patients were discriminated from HS with sensitivity and specificity of 90% and 82% when both plasma cirDNA and csb-cirDNA were analyzed. Discrimination of NSCLS from COPD patients was characterized by 88% sensitivity and 80% specificity. Values of RARB2 IM significantly increased in csb-cirDNA and plasma cirDNA from COPD (23%) and NSCLC (35%) patients compared with HS (11%) ($P<0.05$). RASSF1 IM values of plasma cirDNA and csb-cirDNA did not differ between COPD and HS. RARB2 gene IM increase was associated with advanced stage of NSCLC.

Conclusion: Concentration changes of ACTB and LINE-1 fragments demonstrate a strengthening of the processes in lung cancer leading to unequal representation of the genomic DNA fragments in cirDNA of blood. Epigenetic alterations of tumor suppressor genes in the total cirDNA were found to be associated with lung cancer development and progression. Methylation status of two candidate epigenetic markers (RARB2 and RASSF1A genes) in the cirDNA from plasma and csb-cirDNA fractions in combination was found to be valuable for lung cancer diagnostics and tumor staging.

Acknowledgements: The research has been carried out with support of the grants from RFBR № 11-04-12105-offi-m-2011, SB RAS Program in collaboration with other scientific organizations № 65, RAS Program "Fundamental Science for Medicine" № 23, Federal Special-Purpose Program " Scientific, Academic, and Teaching Staff of Innovative Russia" 2009-2013 (№ P256).

INTEGRATED APPROACH TO MOLECULAR DYNAMICS STUDY OF PROTEINS AND PROTEIN-DNA COMPLEXES

Popov A.V.*, Zharkov D.O., Vorobyov Y.N.
Institute of Chemical Biology and Fundamental Medicine SB RAS, Novosibirsk, Russia
e-mail: apopov@niboch.nsc.ru
** Corresponding author*

Key words: *molecular dynamics, trajectory analysis, base excision repair, glycosylase*

Motivation and Aim: Within the bounds of computer systems designed to predict new pharmacological targets, an integrated approach to modelling and analysis of proteins and DNA-protein complexes should be developed. Molecular dynamics (MD) simulations of biopolymers on meaningful time scales produce large trajectories rarely amenable to manual analysis; therefore a convenient analysis tool is required. The approach should be applied to model and analyze structure and dynamics of some vital enzymes to affirm its suitability and topicality.

Methods and Algorithms: BioPASED, a molecular dynamics modelling program, is a part of BISON complex and uses mathematical model of molecule in the terms of classic molecular mechanics. GUI-BioPASED, a graphical web interface to the BioPASED, was developed to assist in error-free task formation and preliminary structure validation [1]. MDTRA, a molecular dynamics trajectory analyzing program, was developed to facilitate the process of trajectory analyzing, meaningful data extraction and representation. All the tools work in close liaison and form a conveyor from initial PDB structure to numeric data and plot images.

Results: The approach described was applied to analysis of structure and dynamics of DNA-*N*-glycosylases involved in the process of DNA base excision repair. Three different wild-type enzymes were modeled (*Lactococcus lactis* Fpg, *Bacillus stearothermophilus* MutY, human OGG1) with a DNA helix fragment containing a lesion 8-oxoguanine (oG) which is their common substrate. Fpg was modeled in four different charge states, two states of two amino acids of the active site to discover the best charge pair. MutY was modelled with either oG or G opposite to A to study its substrate discrimination. The same was done for OGG1 (oG:C and oG:A pair). Analysis of trajectories shed some light to how these enzymes function. To investigate a mutation impact on structure stability and catalytic activity, three OGG1 mutants were also modeled (C253I, C253L, Q315W). A noticeable difference between these mutants, and between the mutants and the wild type, was shown; it coincided well with the steady state kinetics results.

Conclusion: The integrated approach to molecular dynamics study was developed and proved useful in studies of structure, dynamics and function of some DNA repair enzymes. It can be used to perform high-quality molecular dynamics experiments, including prediction of new pharmacological targets, and analyze their outcome.

Availability: The programs are available on the website (bison.niboch.nsc.ru).

Acknowledgements: This research was supported by the Presidium of the Russian Academy of Sciences (6.14) and by Russian Foundation for Basic Research (11-04-00807, 09-04-00136).

References:

1. A.V. Popov, Yu.N. Vorobjev. (2010) GUI-BioPASED: A Program for Molecular Dynamics Simulations of Biopolymers with a Graphical User Interface, *Molecular Biology (Moscow)*, **44**: 648-654.

NEW ALGORITHM FOR IDENTIFICATION OF INDIVIDUAL DIFFERENCES IN GENE EXPRESSION

Pošćić F. *[1], Khlopova N.S.[2]
[1] *Udine University, Udine, Italy, e-mail: filip.poscic@uniud.it;*
[2] *RSAU-MTAA named after K.A. Timiryazev, Moscow, Russia*
** Corresponding author*

Key words: *probabilistic similarity index, microarray, gene expression, Sus scrofa*

Motivation and Aim: A methodology for identification of individual differences without replicates in gene expression is described. Comparing two or more groups of individual is well known as analysis of variance. However, as the microarray requires a lot of processing and it is quite expensive, it is not always possible to have replicates.

A few articles concerning individual variability used replicates or simply reported genes with 2, 3, 4-fold change criteria as up or down-regulated genes [1]. This method ignores the fact that these differences may or may not account biological effects.

Housekeeping genes (hkgs) are widely used as internal controls in a variety of study types like microarrays and it is possible to choose appropriate internal controls according to procedures such as those from [2].

Methods and Algorithms: As there exist no *a priori* definition of similarity, the use of Goodall probabilistic similarity index using Manhattan distance (PSI) [3] seems to be appropriate. In the PSI an ordering relationship between individuals for each gene is established according with everyday concepts of similarity, but also taking into account the probability that a random sample of two would have the values in question.

Since the PSI procedure is computationally consuming, as an approximation one may divide the distribution into a convenient number of groups. Under the assumption of limited variance of hkgs, we developed a new algorithm for not-arbitrary grouping of individuals. Each individual is situated in a specific group according to its intra-group similarity. Variance of each group should not be larger than those in the hkgs.

For the validation of our model we analysed real data. The study was performed on *Sus scrofa* individuals from the same breed, of the same age and raised in the same sheds. Standard procedures for the microarray technology were applied. For accounting systematic variations occurred during experiments, sample values were normalized [4].

Results: Our results were than compared with those from simple fold change criteria. We clearly demonstrated the power of the model for searching the potentially differently expressed genes in individuals.

Conclusion: The PSI and our model were applied successfully on our data for identification of significantly differently expressed genes between individuals.

Availability: The whole procedure has been programmed in C++ for massive analysis of data on personal computer and is currently on check for optimization. It will be available upon request.

References:

1. Sasaki A. et al. (2003) Individual differences in gene expression in primary cultured renal cortex cells derived from japanese subjects, IPSJ-DC, 2: 710-715.
2. Lee S. et al. (2007) Identification of Novel Universal Housekeeping Genes by Statistical Analysis of Microarray Data, JBMB, 40: 226-231.
3. Goodall D.W. (1966) A New Similarity Index Based on Probability, Biometrics, 22: 882-907.
4. Smyth G.K., Speed T.P. (2003). Normalization of cDNA microarray data. Methods 31, 265-273.

IDENTIFICATION OF BIOLOGICAL TARGETS FOR VIRTUAL SCREENING OF INHIBITORS OF REPLICATION OF TICK-BORNE ENCEPHALITIS VIRUS

Potapov V.V.*[1], Potapova U.V.[1], Belikov S.I.[1], Sidorov I.A.[2], Novopashin A.P. [2], Pozdnyak E.I.[2], Mukha D.V. [3], Feranchuk S.I. [4]

[1] *Limnological Institute, SB RAS, Irkutsk, Russia;*

[2] *Institute of System Dynamics and Control Theory, SB RAS, Irkutsk, Russia;*

[3] *Institute of Bioorganic Chemistry, NASB, Minsk, The Republic of Belarus;*

[4] *Belarusian State University, Minsk, The Republic of Belarus*

e-mail: vpotapov@lin.irk.ru

** Corresponding author*

Key words: *virtual screening, inhibition of serine protease, tick-borne encephalitis virus*

Motivation and Aim: There are technologies in modern pharmacology which radically change the ways of creating drugs. One such technologies is a virtual screening of libraries of potential compounds. The success of this method in the search for ligands that inhibit the replication of flaviviruses confirmed by a number of works. This fact gives a hope in efficiency of the method in the screening of ligands to the tick-borne encephalitis virus (TBEV) proteins.

Methods and Algorithms: The molecular dynamics simulations (MD) were performed using Amber software version 11. The analysis the MD trajectories of two protein structures was performed using the original algorithm [1].

Results: It is known that the substitutions in serine proteases correlate with differences in the pathogenicity of strains of tick-borne encephalitis [2].

It was found that the substitutions lead to a change in the conformation of the protease [1]. There are regions in the protein globule, which are localized near the active site of the protein and change the conformation depending on strain. The blocking of these regions by a ligand could change the activity of the protease. The purpose this work is to identify chemical compounds, that are able to specifically bind and indirectly inhibit the function of virus serine protease. It will be made to eliminate possibility of undesirable blocking of human serine proteases.

*Conclusion:*The target protein was found for high-specific inhibition of replication TBEV.

Acknowledgements: This work was funded in part by the interdisciplinary integration project of SB RAS № 22, Grant ISTC № 4006 .

References:

1. Potapova U.V., Feranchuk S.I., Potapov V.V., Kulakova N.V., Kondratov I.G., Leonova G.N. Belikov S.I. (2012) NS2B/NS3 protease: allosteric effect of mutations associated with the pathogenicity of tick-borne encephalitis virus, *accepted to Journal of Biomolecular Structure and Dynamics*
2. Belikov, S.I., Leonova, G.N., Kondratov, I.G., Romanova, E.V., & Pavlenko, E.V. (2010). Coding nucleotide sequences of tick-borne encephalitis virus strains isolated from human blood without clinical symptoms of infection, *Russian Journal of Genetics,* **46:** 315-322.

A SCREENING OF G-QUADRUPLEX MOTIFS AS A STRUCTURAL BASIS OF APTAMERS TO TICK-BORNE ENCEPHALITIS VIRUS GLYCOPROTEIN

Potapova U.V.*[1], Potapov V.V. [1], Kondratov I.G.[1], Solovarov I.S., Belikov S.I.[1], Vasiliev I.L. [2]

[1] *Limnological Institute, SB RAS, Irkutsk, Russia;*

[2] *Institute of System Dynamics and Control Theory, SB RAS, Irkutsk, Russia*

e-mail: potapova@lin.irk.ru

** Corresponding author*

Key words: *G-quadruplex, aptamer structure, tick-borne encephalitis virus*

Motivation and Aim: A replacement of specific antibodies by DNA-aptamers is a modern trend in a therapy of viral infections. Anti-viral substances for a hepatitis C and other viruses are currently developed. A transfer of aptamers inside a cell to achieve an interaction with target proteins is a main problem of aptamer therapy. A purpose of the research is to obtain aptamers to surface protein E of tick-borne encephalitis virus, as this protein is available for aptamers outside a host cell. From a pool of 171 aptamer sequences, the sequences should be found which could selectively bind to glycoprotein E and prevent a penetration of a virion to a cell. The common structural motifs should be determined in the selected sequences.

Methods and Algorithms: A frequency analysis was used as a method of an aptamer screening. The complete enumeration gave 8 most frequent segments with a length not less than 10 nucleotides. Web-service QGRS Mapper [1] shows motifs of G-quadruplexes in the selected segments.

Results: G-quadruplexes are the structural backbone of aptamers and are quite stable relative to simple structural motifs of hairpins and pseudo-knots. G-quadruplexes are most often in DNA aptamers [2].

The possible tertiary structures of the selected aptamers are predicted from a known tertiary structures of G-quadruplexes.

Conclusion: The results obtained could lead to a development of a safe antiviral medicine on a base of highly specific DNA aptamers.

Acknowledgements: This work was funded in part by the Russian Ministry of Science and Education (State Contract No. 16.512.11.2258), the interdisciplinary integration project of SB RAS № 141.

References:

1. Kikin O, (2006). QGRS Mapper: a web-based server for predicting G-quadruplexes in nucleotide sequences, *Nucleic Acids Research,* **34:** 676–682
2. А.В. Кульбачинский (2006). Методы отбора аптамеров к белковым мишеням, *Успехи биологической химии,* **46:** 193-224

NS2B/NS3 PROTEASE: ANALYSIS OF ALLOSTERIC EFFECTS OF MUTATIONS ASSOCIATED WITH THE PATHOGENICITY OF TICK-BORNE ENCEPHALITIS VIRUS

Potapova U.V.*[1], Potapov V.V. [1], Kondratov I.G.[1], Mukha D.V. [2], Feranchuk S.I. [3], Leonova G.N.[4], Belikov S.I.[1]

[1] *Limnological Institute SB RAS, Irkutsk, Russia;*

[2] *Institute of bioorganic Chemistry, NASB, Minsk, The Republic of Belarus;*

[3] *Belarusian State University, Minsk, The Republic of Belarus;*

[4] *Institute of Epidemiology and Microbiology SB RAMS, Vladivostok, Russia*

e-mail: potapova@lin.irk.ru

** Corresponding author*

Key words: *flaviviral protease, pathogenicity, conformational analysis, molecular dynamic*

Motivation and Aim: The paper consider allosteric effects of amino acids substitutions in a NS3 protein associated with the pathogenicity of tick-borne encephalitis virus (TBEV) [1]. The protease domain of the TBEV NS3 protein has two amino acid substitutions, 16 R→K and 45 S→F, in the highly pathogenic and poorly pathogenic strains of the virus, respectively [1].

Methods and Algorithms: Two models of the NS2B-NS3 protease complex of the TBEV were constructed by homology modelling using the crystal structure of West Nile virus protease as a template. The tertiary structures of the protease were modelled by the Nest program of the Jackal package. The MD simulations were performed using Amber software version 11. 20 ns molecular dynamic simulations were performed for both models, the trajectories of the dynamic simulations were compared, and the mathematical analysis of the trajectories was carried on [2].

Results: It was found that two amino acids substitutions lead to stable change in the conformation of the protease. These differences lead to the spatial distance between the hydrophilic segment of the NS2B protein and catalytic triad of the NS3 viral protease [2].

*Conclusion:*We propose that conformational changes in the protease active, caused by two amino acid substitutions, can influence polyprotein processing and the virulence of the virus.

Acknowledgements: This work was funded in part by the joint projects of BRFFR and SB RAS № 22, Grant ISTC № 4006.

References:

1. Belikov, S.I., Leonova, G.N., Kondratov, I.G., Romanova, E.V., & Pavlenko, E.V. (2010). Coding nucleotide sequences of tick-borne encephalitis virus strains isolated from human blood without clinical symptoms of infection, *Russian Journal of Genetics,* **46:** 315-322.
2. Potapova U.V., Feranchuk S.I., Potapov V.V., Kulakova N.V., Kondratov I.G., Leonova G.N. Belikov S.I. (2012) NS2B/NS3 protease: allosteric effect of mutations associated with the pathogenicity of tick-borne encephalitis virus, *accepted to Journal of Biomolecular Structure and Dynamics*

GENE-CENTRIC KNOWLEDGEBASE ON THE WEB

Poverennaya E.V.*, Bogolyubova N.A., Lisitsa A.V., Ponomarenko E.A.
Institute of Biomedical Chemistry RAMS, Moscow, Russia
e-mail: k.poverennaya@gmail.com
** Corresponding author*

Key words: *knowledgebase; chromosome; Human Proteome Project*

Motivation and Aim: We developed Gene-Centric Knowledgebase, which mission is to represent, store and exchange of proteomics data in a heat-map format, adopted for Chromosome-centric Human Proteome Project [1].

Methods and Algorithms: List of UniProt accession number for the proteins encoded by 18-th chromosome was imported to create the scaffold of the heat-map. For each protein the descriptors were produced from the data resources. The descriptors characterized proteins in the global data resources, including UniProt, NCBI, PRIDE, PeptideAtlas, GPMDB, KEGG, OMIM, PubMed, STRING, PDB and data from papers.

We used Knowledgebase to encapsulate the proprietary experimental data produced in the course of 18-th chr-centric Russian part of Human Proteome Project [2]. Descriptors were allocated to indicate the identified proteins and transcripts. The value of each descriptor was normalized to the median and color-coded. List of proteins with corresponding descriptors is displayed as heat-matrix, navigable through the Web.

Results: Knowledgebase content manager enables to sort the descriptors and to select them for copying into the user sets. With this features it is possible, for instance, to select all the salient proteins frequently observed in mass-spectrometry repositories and then to compare frequencies with abundance of proteins and transcripts in the cell.

Conclusions: Gene-Centric Knowledgebase updates the descriptors through the simple automated programming interface. Knowledgebase can be useful to integrate user-defined data sources within the scope of Human Proteome Project.

Availability: Pilot version is available at www.kb18.ru.

References:

1. Paik YK, Jeong SK, Omenn GS et al, The Chromosome-Centric Human Proteome Project for cataloging proteins encoded in the genome, Nat Biotechnology, 3, 2012, 221-223
2. Archakov A, Aseev A, Bykov V et al, Gene-centric view on the human proteome project: the example of the Russian roadmap for chromosome 18, Proteomics, 11, 2011, 1853-1856.

THE TECHNIQUES AND TOOLS FOR THE SOLVING BIONFORMATICS TASKS IN THE DISTRIBUTED COMPUTING SYSTEMS

Pozdnyak E.I.*[1], Oparin G.A.[1], Novopashin A.P. [1], Sidorov I.A.[1], Potapov V.V. [2], Potapova U.V. [2], Belikov S.I. [2], Mukha D.V. [3], Feranchuk S.I.[4]
[1] *Institute for System Dynamics and Control Theory SB RAS, Irkutsk, Russia;*
[2] *Limnological Institute SB RAS, Irkutsk, Russia;*
[3] *Institute of bioorganic Chemistry, NASB, Minsk, The Republic of Belarus;*
[4] *Belarusian State University, Minsk, The Republic of Belarus*
e-mail: pozdnyak@icc.ru
** Corresponding author*

Key words: *distributed computing, related tasks, distributed packages of applied programs*

Motivation and Aim: There are a lot of program tools for describing and solving bioinformatics tasks in the distributed computing systems. Some of these tools allow to describe a set of related tasks which contain many independent subtasks which can be started in the parallel mode on the clusters. This technique permits to significantly speeding up the computation for various bioinformatics programs.

Methods and Algorithms: The most popular form for the description of related tasks is directed acyclic graph (DAG). This form describes the dependence between input parameters of one program and output parameters of another program.

Results: Based on the DAG technique was developed a DISCOMP toolkit [1] which allows us to develop a distributed packages of applied programs. By using DISCOMP toolkit was developed a parallel scheme of an algorithm generalization of the taxonomic classifier 'CARMA' [2] (Figure 1).

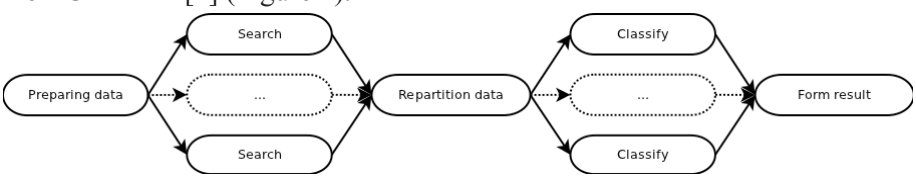


Figure 1. Algorithm of program package CARMA.

Conclusion: DISCOMP toolkit oriented to automating a generation of chains of related bioinformatics programs and run it on the clusters in the parallel mode.

Acknowledgements: This work was funded in part by the interdisciplinary integration project of SB RAS № 22, Grant ISTC № 4006.

References:

1. Oparin G.A., Feoktistov A.G., Sidorov I.A. (2009). Technology of the organization distributed packages of applied programs in the DISCOMP toolkit, Modern technologies. System analysis. Modelling, 2: 175-180.
2. Pozdnyak E.I., Sidorov I.A., Galachyants Y.P. (2011). Algorithm generalization of the 'CARMA' taxonomic classifier, Vestnik ISTU, 9: 11-15.

MECHANISMS OF AMPA RECEPTOR TRAFFICKING AS A BASE OF CHANGING THE SYNAPTIC EFFICIENCY

Proskura A.L.*[1], Malakhin I.A.[1, 2], Zapara T.A.[1]
[1] *Design Technological Institute of Digital Techniques SB RAS, Novosibirsk, Russia;*
[2] *Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia*
e-mail: annleop@mail.ru
** Corresponding author*

Key words: *LTP, vesicle transport, receptor recycling, synaptic plasticity*

Motivation and Aim: Synaptic plasticity is mediated by the receptor density change mechanisms in hippocampus. AMPAR [AMPA (alpha-amino-3-hydroxy-5-methylisoxazole-4-propionic acid) receptor] are synthesized in neuronal soma and delivered to synapses on the different stages of endocytic pathway along variously cytoskeleton structures. Dendritic spines are important sites for AMPAR trafficking, they contain the basic components of endosomal recycling. On the base of analysis and integration of recent experimental data we had developed the model of subcellular processes that describes the mechanisms of AMPAR changing and maintenance on synaptic membrane.

Methods and Algorithms: To reconstruct the model of molecular network the GeneNet technology was used.

Results: Using GeneNet technology, we reconstructed the graphic model reflecting the main events of AMPAR delivery through various stages of vesicle trafficking. The analysis of reconstructed spine interactome permitted to extract the key processes that control AMPAR density on synaptic membrane. Biosynthesis pathway, including synthesis and subunit assembly of receptors within endoplasmic reticulum (ER), exit from ER to Golgi, exit from trans-Golgi net, mediates the delivery of newly synthesized receptors into a synapse. The speed of exit from ER is defined by subunit structure of a receptor. The transport of AMPAR into synapses can occurs in two different ways. GluR2–GluR3 heterotetramers circulates continuously into and out of synapses, whereas GluR1-containing receptors enter into synapses in an activity-dependent manner during long-term potentiation (LTP). Various proteins from the small GTPase family are the main regulators of the processes described above. SAR1b, ARF1, ARF6, Dynamin, particularly, are responsible for correct assembly and detachment of vesicles transporting the AMPAR. GTPases (Rab4, Rab11, Rab8, Rab5) from subfamily Rab regulates the processes of sorting and interacting the vesicles with protein-motors that mediate the moving of vesicles along microtubules or actin filaments. In our network some mechanisms that coordinate vesicle assembly and actin remodelling during exo- and endocytosis of AMPAR are presented.

Conclusion: The state of receptors of synapse defines switching the mechanisms that control the density of AMPAR. During LTP expression GluR1-AMPAR are actively delivered in synaptic zone, while the mechanisms of constitutive recycling of GluR2–GluR3-AMPAR maintain the new receptor density.

Availability: available on request from the authors.

Acknowledgements: The work was supported by RAS base fundamental research project VI.53.1.3, Integration project presidium SB RAS № 136, RFBR grant № 12-01-00639.

STUDY OF CONFORMATIONAL FLEXIBILITY OF *E. COLI* RNA POLYMERASE ALPHA SUBUNIT INTERDOMAIN LINKER

Purtov Yu.A.*, Kondratyev M.S., Ozoline O.N., Komarov V.M.
Institute of Cell Biophysics RAS, Pushchino, Moscow region, Russian Federation
e-mail: yapurtov@yahoo.com
** Corresponding author*

Key words: *alpha-subunit RNA-polymerase, interdomain linker, structure, dynamics*

Motivation and Aim: In spite of the fact that process of transcription initiation in prokaryotes is well understood some of its aspects are complicated for experimental study. In particular it can be attributed to positioning of RNA polymerase alpha subunits on DNA. This protein consists of two domains (N- and C-terminal) joined with linker preventing protein crystallization and determination of its spatial structure. Earlier it was shown that changes in amino acid sequence of linker led to decrease of promoter recognition efficiency [1]. This work is dedicated to study of structural and dynamical aspects of interdomain linker flexibility by use of computational full-atomic modeling.

Methods and algorithms: homology analysis, sequence alignment, quantum-chemical PM3 method (MOPAC2009), molecular dynamics (GROMACS), molecular docking (HEX, Autodock Vina).

Results: analysis of spatial structure and dynamic flexibility of *E.coli* RNA polymerase alpha subunit linker as well as its artificially selected homologues allowed to determine significant role of several amino acid residues in linker topology. We assume that charged amino acids and prolines to be essential for this process. Charged residues (Arg, Lys, Asp, Glu) affect the linker structure forming salt bridges, while prolines provide appropriate bending and strength of peptide chain. Molecular docking of linker on DNA showed its potential capability to interact with both major and minor DNA groove. Comparison of amino acid sequence of linker and its homologues from other prokaryotes allowed to determine natural variability of this region and suggest a hypothesis that linker composition is dependent on ecology of bacterial species.

Conclusion: investigation of structure of *E.coli* RNA polymerase alpha subunit interdomain linker gives a possibility to state that amino acid sequence of this protein region directly affects alpha subunit C-terminal domain positioning on recognized DNA sequence. We consider this linker flexibility preventing alpha subunit crystallization as adapting mechanism for interaction with variety of promoter sequences and transcription factors.

References:

1. Fujita N, Endo S, Ishihama A. (2000) Structural requirements for the interdomain linker of alpha subunit of Escherichia coli RNA polymerase. Biochemistry. 39, 6243-6249.

COMPARATIVE ANALYSIS OF TRIPLETS FREQUENCY IN MITOCHONDRIAL GENOMES

Putintseva Yu.A.[1, 2]
[1] *Siberian Federal University, Krasnoyarsk, Russia;*
[2] *Krasnoyarsk state medical university, Krasnoyarsk, Russia*
e-mail: yuliya-putintseva@rambler.ru

Key words: *triplet frequency, codons, elastic maps*

Motivation and Aim: A study of statistical properties of nucleotide sequences may bring a lot towards the architecture of genome as well as about the relation between structure and function encoded in these former. A consistent and comprehensive investigation of the features and peculiarities is based on the study of frequency dictionary of a nucleotide sequence. Such approach answers the questions concerning the statistical and information properties of DNA sequences. A frequency dictionary, whatever one understands for it, is rather multidimensional entity.

Methods and Algorithms: 2461 mitochondrion genomes were retrieved from the page of European Bioinformatics Institute (http://www.ebi.ac.uk/genomes/organelle.html). The list of available genomes is inhomogeneous, from the point of view of the equity of the number of species of various genders enlisted into the database. To eliminate the effect of the possible bias associated with this heterogeneity, we hashed the databases: a single genome from a gender was selected randomly, while the other ones were eliminated from the database. It resulted in a decrease of the number of entries in the database up to 1651 ones.

A standard unsupervised classification technique was implemented to develop a classification of the genomes in 63-dimensional space of triplets and codon frequencies. We used ViDaExpert software [1] to do that.

Results: 1651 mitochondrial genomes were classified in 63-dimensional space of triplets and codon frequencies for four genes with the K-means method and method of elastic maps. Picture of the distribution of genomes in this space is now being studied. However, it is clear that the picture is considerably different for different genes and is not random.

Conclusion: Genetically, the mitochondrion genomes are rather conservative, thus providing a good raw for knowledge extraction. The distribution of mitochondrion genes in 63-dimensional space of triplets and codon frequencies is far from a random one.

Acknowledgements: This work was supported by grant from Russian government department of Science and Education to Siberian Federal University «The genetic researches of the Siberian larch».

References:

1. 1. http://bioinfo-out.curie.fr/projects/vidaexpert/

GENETIC SUSCEPTIBILITY PROFILE FOR COMORBIDITY VARIANTS OF MULTIFACTORIAL DISEASES

Puzyrev V.P.*[1], Makeeva O.A.[1], Barbarash O.L.[2], Sleptcov A.A.[1], Markova V.V.[1], Polovkova O.G.[1]

[1] *Research Institute of Medical Genetics SB RAMS, Tomsk, Russia;*
[2] *Research Institute of Complex Issues of Cardiovascular Diseases SB RAMS, Kemerovo, Russia;*
[3] *Genoanalytica, LLC*
e-mail: valery.puzyrev@medgenetics.ru
** Corresponding author*

Key words: *multifactorial diseases, genetic susceptibility, cardiovascular continuum, syntropy*

Motivation and Aim: The "Cardiovascular Continuum" (CVC) was described in 1991 by Dzau and Braunwald to explain progression of coronary heart disease through other complications and diseases to inevitable end stage of heart failure. Concept of CVC includes several diseases such as coronary artery disease (CAD), arterial hypertension (AH), metabolic syndrome, and diabetes mellitus type 2 (DM2) (Dzau, Braunwald, 1991). In 1921 Pfaundler and von Seht used the term 'syntropy' to designate diseases, which tend to co-occur with each other in patients (or in families) more often than it could be expected by chance (Pfaundler, von Seht, 1921). Based on these two concepts, the term "syntropy genes" (SG) was proposed to designate a set of functionally interacting, co-regulated genes involved in common biochemical and physiological pathways leading to syntopy (Puzyrev, 2008). Thus, the aim was to explore the genetic profile of CVC and to identify SG.

Methods and Algorithms: A large sample of patients with ischemic heart disease and population sample were analyzed to select subgroups for present study. A total of 309 patients out of 800 were selected according to the following criteria: "syntropy" subgroup diagnosed with CAD, AH, DM2, and dyslipidemia in each patient (N=68); comorbidity of CAD and AH with other cardiovascular pathology excluded (n=180), and a subgroup of patients with CAD only (other diseases excluded, N=61). Healthy subjects with normal cardiovascular endophenotypes (N=131) were selected out of a sample of 1600 individuals. Genotyping was done using Illumina Human custom chip microarrays with a panel of markers used for direct-to-consumer genomic service "My Gene" (www.i-gene.ru) (Genoanalytica, LLC). For statistical analysis R v2.14.0 software environment was used, including specialized packages "GenABEL", "snpStats" and "genetics". Predictive value of each candidate SNP was tested using AUC (area under curve).

Results and conclusion: Syntropy group differed significantly from other samples analyzed. Pathway of *ITGA4, KLF7,* and *TAS2R38* genes was involved in the development of this comorbidity. Advanced classifier analysis yielded that *KLF7* rs7568369 reached AUC of 63% and three other SNPs *(LDLR* rs2738446, rs688*,* and *CDKN2A* rs1333048) reached maxAUC of 63%. GG genotype of the rs6501455 located in a region between *KCNJ2* and *SOX9* genes yielded most substantial risk effect in reference to CVC syntropy (OR 3,91; 95% CI 1,56-10,33; P<0.0016). GG genotype of the rs7568369 in *KLF7* had highest protective effect in reference to CVC syntropy (OR 0.34; 95% CI 0,16-0,68; $P<7*10^{-4}$). Cluster analysis which involved 90 SNPs, related to different cardiovascular phenotypes, showed that syntopy forms a separate cluster, while other subgroups are close to each other in genetic characteristics.

The study demonstrates that CVC syntropy differs significantly in genetic characteristics from other forms of cardiovascular parthology and has specific genes (SG) involved.

CLUSTER ANALYSIS OF SIGNIFICANT REGULATORS AS NEW APPROACH TO PATIENTS SUBTYPING

Pyatnitskiy M.*[1, 2], Mazo I.[1, 3], Daraselia N.[3], Shkrob M.[3], Kotelnikova E.[1, 4]
[1] *Ariadne Diagnostics LLC, Rockville, USA;*
[2] *Institute of Biomedical Chemistry, RAMS, Moscow, Russia;*
[3] *Reed Elsevier, Amsterdam, Netherlands;*
[4] *Institute for Information Transmission Problems, RAS, Moscow, Russia*
e-mail: mpyat@ariadne.net
** Corresponding author*

Key words: *sub-network enrichment analysis, gene expression, cluster analysis*

Motivation and Aim: Personalized approach to medical treatment is one of the most important challenges in the modern medicine. To address this problem it is necessary to reveal different mechanisms within the same disease that would distinguish patients from several subgroups. Standard approaches to patient clustering using gene expression often perform poorly due to overfitting which frequently takes place in case of tens of thousands of features. Sub-Network Enrichment Analysis (SNEA) is a topology-based variant of gene set enrichment analysis and allows projecting differential profiles of thousands of genes onto much fewer key significant regulators [1]. Here we propose combined approach which includes consecutive usage of SNEA and biclustering of identified regulators and samples.

Methods and Algorithms: We applied SNEA separately to each sample differential profile. Next we clustered obtained regulators into groups and calculated statistic reflecting activity of each cluster. Number of regulator clusters was determined automatically using silhouettes. Samples were clustered according to values of statistic and correspondence to ground-truth labels was estimated via Rand index. The whole pipeline was cross-validated to ensure overall stability of the algorithm. SNEA was performed with Pathway Studio 9.0 from Ariadne Genomics, all other analysis were done with set of in-house R scripts.

Results: We applied our approach to several expression datasets including studies on nevus/melanoma differentiation and colon cancer. The same datasets were clustered using well-known PAM method and we found that our method in most cases performs better. We also analyzed microarray data on patient's response to anti-EGFR therapy with cetuximab and were able to identify group of non-responders. The most prominent mechanism was dependent on TGFb-SMAD pathway and epithelial-to-mesenchymal transition.

Conclusion: We showed that proposed approach performs well for clustering patients using gene expression data. The main benefit of our method is that obtained clusters of regulators provide valuable insight into the molecular mechanisms and pathways closely related to a clinical outcome for individual patient, thus serving as biologically meaningful feature selection. Also genes downstream of identified significant regulators may also serve as candidate biomarkers for discriminating between various conditions.

Availability: Source R code for processing SNEA results is freely available from authors.

References:

1. A.Y.Sivachenko et al. (2007) Molecular networks in microarray analysis, *Journal of bioinformatics and computational biology,* **5**(2B): 429-56.

NOVEL APPROACH TO META-ANALYSIS OF MICROARRAY DATASETS FOR IDENTIFICATION OF NEW BIOMARKERS AND POTENTIAL DRUG TARGETS

Pyatnitskiy M.*[1, 2], Kotelnikova E.[1, 3], Shkrob M.[4], Ferlini A.[5], Daraselia N.[4], Mazo I.[1, 4], Schwartz E. [1]

[1] *Ariadne Diagnostics LLC, Rockville, USA;*
[2] *Institute of Biomedical Chemistry, RAMS, Moscow, Russia;*
[3] *Institute for Information Transmission Problems, RAS, Moscow, Russia;*
[4] *Reed Elsevier, Amsterdam, Netherlands;*
[5] *Department of Experimental and Diagnostic Medicine, University of Ferrara, Ferrara, Italy*
e-mail: mpyat@ariadne.net
** Corresponding author*

Key words: *subnetwork enrichment analysis, biomarkers, gene expression, meta-analysis*

Motivation and Aim: Analysis of high-throughput gene expression data has become an area of intensive research and has already proven it's utility. However, finding new biomarkers and potential drug targets from expression profiling still remains a challenge due to the limited number of available samples and lack of overlap to accomplish cross-dataset comparisons. Here we propose a novel computational approach for drug target and biomarker discovery using a comprehensive meta-analysis of multiple gene expression datasets.

Methods and Algorithms: We accessed 5 publicly available human muscle gene expression datasets related to Duchenne muscular dystrophy (DMD). To ensure overall reproducibility we constructed additional datasets by aggregating 4 out of the 5 datasets in a method similar to leave-one-out cross-validation. Each dataset was processed with Sub-Network Enrichment Analysis (SNEA), a topology-based variant of gene set enrichment analysis using a global literature-extracted expression regulation network. SNEA projects differential profiles of thousands of genes onto much fewer key significant regulators. We also identified consistently differentially expressed genes by arranging a selection of different ranking statistics. SNEA was performed with Pathway Studio 7.1 from Ariadne Genomics, all other analyses were done with a set of in-house scripts for R/BioConductor.

Results: We discovered a disturbance in the activity of several muscle-related transcription factors, regulators of inflammation, regeneration, and fibrosis. Almost all SNEA-identifed regulators of down-regulated genes corresponded to a single common pathway important for fast-to-slow twitch muscle fiber type transition. We hypothesize that this process can affect the severity of DMD symptoms, making corresponding regulators and downstream genes valuable candidates as potential drug targets and exploratory biomarkers.

Conclusion: Using DMD as an example, we have demonstrated the possibility to decipher regulatory mechanisms of disease along with corresponding exploratory biomarkers on the basis of meta-analysis of multiple microarray datasets. A number of the predicted expression regulators are previously known to be involved in DMD, suggesting that the others will also be verified hereafter. This means that all of the proposed regulators can be considered for further drug discovery, whereas their consistently differentially expressed downstream genes may serve as exploratory biomarkers.

Availability: Source R code is freely available from authors.

BIOINFORMATIC SEARCH AND PHYLOGENETIC ANALYSIS OF THE PLANT-SPECIFIC MAPS IN GENOMES OF MONOCOTS AND DICOTS

Pydiura N.A.*, Karpov P.A., Blume Ya.B.
Institute of Food Biotechnology and Genomics NASU, Kyiv, Ukraine
e-mail: nikolay.pydiura@gmail.com
** Corresponding author*

Key words: *MAP-60, MAP-70, EB1, bioinformatics search, phylogeny*

Motivation and Aim: The organization of plant microtubules (MTs) requires a diverse composition of plant-specific microtubule-associated proteins (MAPs) as they lack centrioles that organize MTs in the cytoplasm of animal cells. Affecting microtubule behavior, MAPs, are important targets of biotechnology. However, the sequences of plant-specific MAPs are identified and annotated only in several plant genomes. Previously [1] we have performed the bioinformatic search for plant homologues of animal structural MAPs in *A. thaliana* genome. In this study we perform the bioinformatics search and consequent multiple alignment and phylogenetic analysis of the plant-specific MAPs of the families MAP60 and MAP70 and of the EB1 family in the genomes of dicots (*Arabidopsis, Brassica, Glycine, Medicago*) and monocots (*Oryza, Hordeum, Triticum, Zea*).

Methods and Algorithms: The search of candidate sequences in genomes of selected plants was performed with tblastn tool *vs* the "Nucleotide collection" database, using annotated sequences of *A. thaliana* and *O. sativa* as input. The multiple alignment, alignment curation, construction, visualization and analysis of phylogenetic trees was performed by the Phylogeny.fr (www.phylogeny.fr) all-in-one web service.

Results: We have identified the homologues of the plant-specific protein MAP65-1 in all the selected genus of monocots and dicots, but *Brassica* and homologues of the MAP70-1 and EB1 in all the dicots and monocots, but *Triticum*. Although there were found several short fragments up to 150 amino acids (450b.p.) long, homologous to different fragments of MAP65-1 and MAP70-1 in *Brassica* and *Triticum* genomic sequences, fragments corresponding to the entire sequence of MAP65-1 and MAP70-1 were not found. Thus, *Brassica* and *Triticum* genus genomic sequences require further genomic investigations.

In phylogenetic analysis in all the three cases (for MAP65-1, MAP70-1 and EB1) the sequences of MAPs of monocots and dicots were grouped into two different branches. Moreover, as it was suggested by MAP70-1 and EB1 sequences phylogenetic analysis, the dicots were grouped into two different branches with the branch containing *Arabidopsis, Brassica* being the most divergent from two other dicots (*Glycine* and *Medicago*) and monocots. The most divergent monocot genus is *Zea*.

Conclusion: The differences in amino acid sequences of MAP65-1, MAP70-1 and EB1, suggests that they may interact with the same biotech drugs in different ways.

References:

1. P.A. Karpov, Y.B. Blume (2008) Bioinformatic search for plant homologues of animal structural MAPs in the *Arabidopsis thaliana* genome, In: *The Plant Cytoskeleton: a Key Tool for Agro-Biotechnology.* Blume YB *et al.*, 373-397 (Springer).

IN SILICO STUDIES OF POTENTIAL PHOSPHORESIDUES IN THE HUMAN NUCLEOPHOSMIN/B23: ITS KINASES AND RELATED BIOLOGICAL PROCESSES

Gioser Ramos-Echazábal[1]*, Glay Chinea[2], Rossana Garcia-Fernández[3], Tirso Pons[4]

[1] Department of Animal and Human Biology, Faculty of Biology, University of Havana, Havana 10400, Cuba;
[2] Biomedical Division, Center for Genetic Engineering and Biotechnology, P.O. Box 6162, Havana 10600, Cuba;
[3] Center for Protein Studies, Faculty of Biology, University of Havana, Havana 10400, Cuba;
[4] Structural Biology and Biocomputing Programme, Spanish National Cancer, Research Centre (CNIO), C/Melchor Fernőndez Almagro 3, Madrid E-28029, Spain
e-mail: gio@fbio.uh.cu, gioser.echazabal@gmail.com
** Corresponding author*

Key words: *phosphorylation; prediction; kinases; protein–protein interactions; cellular localization; signaling; pathway; conservation*

Motivation and Aim: Human nucleophosmin/B23 (32 kDa /pI 5.1) is a phosphoprotein involved in ribosome biogenesis, centrosome duplication, and apoptosis. Its function, localization, and mobility within cells, are highly regulated by phosphorylation events (1). Up to 21 phosphosites of B23 have been experimentally verified even though the corresponding kinase is known only for seven of them (Phosida code: P06748 and Phospho.ELM code: P06748). In this work, we predict the phosphorylation sites in human B23 using seven public servers.

Methods and Algorithms: Of these, six were kinase-specific servers (KinasePhos 2.0, PredPhospho, NetPhosK 1.0, PKC Scan, pkaPS, and MetaPredPS) and one was not (DISPHOS 1.3). The results were integrated with information regarding 3D structure and residue conservation of B23, as well as cellular localizations, cellular processes, signaling pathways and protein–protein interaction networks involving both B23 and each predicted kinase.

Results: Thus, all 40 potential phosphosites of B23 were predicted with significant score (>0.50) as substrates of at least one of 38 kinases. Thirteen of these residues are newly proposed showing high probability of phosphorylation considering their solvent accessibility. Our results also suggest that the enzymes CDKs, PKC, CK2, PLK1, and PKA could phosphorylate B23 at higher number of sites than those previously reported. Furthermore, PDK, GSK3, ATM, MAPK, PKB, and CHK1 could mediate multisite phosphorylation of B23, although they have not been verified as kinases for this protein.

Conclusion: Finally, we suggest that B23 phosphorylation is related to cellular processes such as apoptosis, cell survival, cell proliferation, and response to DNA damage stimulus, in which these kinases are involved. These predictions could contribute to a better understanding, as well as addressing further experimental studies, of B23 phosphorylation.

References:

1. Okuwaki M. The structure and functions of NPM1/Nucleophsmin/B23, a multifunctional nucleolar acidic protein. J Biochem 2008;143:441-448.
2. Gnad F, Gunawardena J, Mann M. PHOSIDA 2011: the posttranslational modification database. Nucleic Acids Res 2011;39:D253-D260.
3. Dinkel H, Chica C, Via A, Gould CM, Jensen LJ, Gibson TJ, Diella F. Phospho.ELM: a database of phosphorylation sites--update 2011. Nucleic Acids Res 2011;39:D261-D267.

CLASSIFICATION OF PURIFIED BONE MARROW POPULATIONS SORTED VIA MULTICOLOR FLOW CYTOMETRY, APPLICATIONS IN ACCUTE MYELOID LEUKEMIA

Rapin N.[1, 2], Jendholm J.[1, 2], Theilgaard K.[1, 2], Winther O.[3], Bullinger L.[4], Porse B.T.*[1, 2]

[1] *Biotech Research and Innovation Centre (BRIC), University of Copenhagen, Denmark;*
[2] *Finsen Laboratory / Rigshospitalet, University of Copenhagen, Denmark;*
[3] *Informatics and Mathematical Modelling, Technical University of Denmark (DTU), Denmark;*
[4] *Department of Internal Medicine III, University Hospital of Ulm, Ulm, Germany*
e-mail: bo.porse@finsenlab.dk
** Corresponding author*

Key words: *Acute Myeloid Leukemia, Gene Expression profiling, classification, Survival studies, Clinical studies*

Motivation and Aim: Acute myeloid leukemia (AML) represents a heterogeneous disease that is based on chromosomal and molecular genetic aberration which can been subdivided into distinct prognostic subclasses. AML originates from different leukemic stem cells (LSC) that in turn are descents of normal hematopoietic stem or hematopoietic progenitor cells (HSCs/HPCs). Our hypothesis is that it is not pertinent to identify subclasses and deregulated genes/pathways using genomic strategies by comparison of AML subclasses with each other but rather by comparison with their normal counterparts. A such strategy is not only predicted to subtract most of the 'normal activity' that defines the differentiation stage of the AML population and its normal counterpart but is also likely to define those genes and pathways that are deregulated in AML and thus might contribute to malignant transformation and ultimately to a distinct AML phenotype.

Methods and Algorithms: We have therefore worked on establishing a classifier using microarray expression profiles from purified bone marrow HSCs and myeloid HPCs populations that are publically available or were purified by multicolor flow cytometry cell sorting.

Results: Using this classifier, we are able to map AML microarray expression profiles to their closest normal counterparts. Once this link is established, we compute a set of genes that are up- and down-regulated in the AML vs normal counterpart and demonstrate that these gene sets allow for classification and prognostication of AML patients. The transformed data allows us to classify AML subtypes with a superior accuracy as the raw expression data. We have also applied the transformation to a dataset of "hard to stratify" patients with AML with normal karyotype and found a discriminative behavior regarding survival. We were able to identify a group with worse outcome.

Conclusion: Our findings demonstrated the potential of genomic strategies comparing cancer vs their normal counterparts with respect to classification and prognostication.

References:

1. Haferlach, T. *et al*. 2010, May. *J Clin Oncol* 28(15), 2529–37.
2. Kvinlaug, B. T. *et. al*. 2011, Apr *Cancer Res*.
3. Ishwaran H. et. al. 2010. Random survival forests for competing risks.

POPULATION GENETIC ANALYSIS OF CASPIAN STURGEONS (*ACIPENCER GUELDENSTAEDTII, ACIPENCER PERSICUS*) USING NEXT GENERATION SEQUENCING AND CUSTOMIZED ILLUMINA GOLDENGATE GENOTYPING ASSAY

Rastorguev S.M.*[1], Nedoluzhko A.V.[1], Mazur A.M.[1], Gruzdeva N.M.[1], Tsygankova S.V.[1], Boulygina E.S. [1], Barmintseva A.E.[2], Mugue N.S.[2], Prokhortchouk E.B. [1]

[1] *National Research Centre "Kurchatov Institute" (NRC "Kurchatov Institute"), Moscow, Russia;*

[2] *Russian Federal Research Institute of Fishery and Oceanography (VNIRO), Moscow, Russia*

e-mail: rastorgueff@gmail.com

** Corresponding author*

Key words: *Acipencer gueldenstaedtii, Acipencer persicus, population genetics, Next Generation Sequencing, SNP, GoldenGate*

Motivation and Aim: The population structure and taxonomic status of *A. gueldenstaedtii* and *A. persicus* require to be investigated for determining quotas of catching and to identify biological characteristics of these objects such as homing, migration etc. Moreover, that is interesting to analyze inheritance of molecular markers in objects such as tetraploid sturgeons.

Methods and Algorithms: The transcriptome sequencing of *A. gueldenstaedtii* was done using SOLiD3 system, and polymorphic loci were determined. 384 SNPs were selected using these data to make customized Illumina GoldenGate DNA microarray for genotyping of 96 samples. Then the sturgeons from different locations of Caspian Sea were genotyped using GoldenGate assay and analyzed with several software packages. At the same assay a family group of sturgeon – male, female and their offspring - were analyzed to determine inheritance of examined polymorphisms.

Results: Population structure of Caspian sturgeons was investigated, and molecular markers for species differentiation were found. Differences between southern and northern populations of *A. gueldenstaedtii* and *A. persicus* were observed. A part of examined SNPs were combined into linkage groups.

Conclusion: Next Generation Sequencing technology is a good approach for large scale molecular markers identification, which can be used in population genetics of even non-model species. These markers allow distinguishing populations with a high statistical support. The customized microarray assay is an easy way to perform large scale genotyping.

ASYMMETRICALLY SELF-UPREGULATED (ASSURE) BIOMOLECULAR SYSTEMS

Ratushny A.V., Saleem R.A., Sitko K., Ramsey S.A., Aitchison J.D.*
Institute for Systems Biology, Seattle, WA, USA;
Seattle Biomedical Research Institute, Seattle, WA, USA
e-mail: John.Aitchison@systemsbiology.org
** Corresponding author*

Key words: *regulatory network motif, positive feedback, mathematical modeling, systems biology, robustness*

Motivation and Aim: Biological regulatory networks are composed of numerous motifs, which once recognized and their dynamics understood, can enable prediction of the system behavior. We noted that many positive feedback loops are typified by two regulatory factors that cooperate to activate numerous target genes and feed back to upregulate only one of the regulatory factors themselves. We term this ASymmetric Self-UpREgulation (ASSURE). The prevalence of the ASSURE motif indicates evolutionary selection and we therefore theoretically and experimentally characterized its dynamical properties.

Methods and Algorithms: The symmetric and asymmetric self-upregulated biomolecular systems were examined using mathematical modeling, sensitivity/robustness analysis and bifurcation analysis. The model predictions were experimentally validated using wild-type and engineered *S. cerevisiae* strains, qPCR, semi-quantitative western blot analysis, FACS analysis and competitive growth assay analysis.

Results: Mathematical modeling revealed that asymmetry in a positive feedback network robustly increases the system's responsiveness and, at the same time, allows the system to precisely control the response. To experimentally validate the role of the ASSURE motif we compared the responses of wild-type (asymmetric feedback) and engineered (symmetric feedback) versions of the fatty-acid-responsive gene regulatory network in budding yeast. The measured expression kinetics *in vivo* were consistent with the model predictions and growth analysis revealed that the ASSURE motif confers a fitness advantage.

Conclusion: We systematically explored and revealed key transient and steady-state features of the asymmetrically self-upregulated network motif, at least partially explaining its evolutionary conservation and prevalence in numerous biological systems.

Availability: Available upon request.

Acknowledgements

This study was supported by NIH/NIGMS (R01-GM075152, U54-2U54RR022220 and P50-GM076547).

INTERPLAY OF GENE EXPRESSION NOISE AND ULTRASENSITIVE DYNAMICS AFFECTS BACTERIAL OPERON ORGANIZATION

Ray J.C.J. [1, 2], Igoshin O.A.*[1]
[1] *Department of Bioengineering, Rice University, Houston, Texas;*
[2] *Department of Systems Biology, UT MD Anderson Cancer Center, Houston, Texas*
e-mail: igoshin@rice.edu
** Presenting author*

Key words: *operon, cotranscription, intrinsic noise, correlation, zero-order ultrasensitivity*

Motivation and Aim: Bacterial chromosomes are organized into polycistronic cotranscribed operons, but evolutionary pressures maintaining them are unclear. We hypothesized that operons alter gene expression noise characteristics, resulting in selection for or against maintaining operons depending on network architecture.

Methods and Algorithms: To test this hypothesis we employed numerical simulations with stochastic simulation algorithms and analytical theory with linear noise approximation to assess the effects of operons for network modules representing different functional modes of protein interactions. To test our modeling predictions we employed bioinformatic analysis of *E. coli* chromosome to compare frequencies of noise-minimizing operon architectures compared with randomized controls and to compare the frequencies of operon coupling as a function of protein/mRNA abundance.

Results: Mathematical models for 6 functional classes of network modules showed that three classes exhibited decreased noise and 3 exhibited increased noise with same-operon cotranscription of interacting proteins. Noise reduction was often associated with a decreased chance of reaching an ultrasensitive threshold. Stochastic simulations of the *lac* operon demonstrated that the predicted effects of transcriptional coupling hold for a complex network module. Our bioinformatic analysis found overrepresentation of noise-minimizing operon organization compared with randomized controls. Among constitutively expressed physically interacting protein pairs, higher coupling frequencies appeared at lower expression levels, where noise effects are expected to be dominant.

Conclusion: Our results thereby suggest a central role for gene expression noise, in many cases interacting with an ultrasensitive switch, in maintaining operons in bacterial chromosomes.

Acknowledgements This work was supported by a fellowship from the NLM Computational Biology and Medicine Training Program of the Keck Center of the Gulf Coast Consortia (NIH Grant No. 5 T15 LM007093-16 to JCJR), National Institutes of Health grant 1R01GM096189-01 to OAI funded through joint NSF DMS/NIH NIGMS Mathematical Biology Initiative.

GENOME SEQUENCES OF CENTENARIANS PRODUCE A BASIS FOR GENOME SCALE LONGEVITY STUDIES

Reshetov D.A.[1,2], Shagam L.I.[1,2], Tyazhelova T.V.[1], Grigorenko A.P.[1,3,4], Andreeva T.A.[1,3], Mikhaylichenko O.A.[1,2], Protasova M.S. [1], Goltsov A.Y. [1,3], Zenin A.A.[5], Gusev F.E.[1], Rogaev E.I.*[1,2,3,4]

[1] Vavilov Institute of General Genetics, Moscow; [2] Lomonosov Moscow State University; [3] Research Center of Mental Health, Russian Academy of Medical Sciences, Moscow, Russia; [4] Brudnick Neuropsychiatric Research Institute, Department of Psychiatry, University of Massachusetts Medical School, 01604, MA, USA; [5] Institute of Functional Nuclear Electronics NRNU MEPHI

e-mail: rogaev@vigg.ru, EVGENY.ROGAEV@umassmed.edu

** Corresponding author*

Key words: *centenarians, longevity, whole-genome sequencing*

Motivation and Aim: Human longevity is known to run strongly in families. Its heritability estimates in twin studies range from 23% to 33%. To date, however, little progress yet has been made in identification of genes for longevity using the common molecular-genetic methods such as genetic association of longevity trait with variations in candidate genes or genome wide association studies (GWAS). Identification of specific allelic gene variations contributing to human variations in lifespan can predict presumable therapeutic targets for common diseases associated with aging.

As alternative approach to GWA, recently emerged direct deep sequencing of genome potentially can provide a new insight on biology of longevity. For example, we hypothesized that very rare genetic variations or mutations rather than combinations of common single nucleotide polymorphism (SNP) (tested in GWAS), may underlie exceptional longevity.

Methods and Algorithms: We have sequenced complete genomes of three centenarian individuals (100 years old and older) of Russian origin and three middle-age individuals with ~ 30-40 fold coverage using Illumina's HiSeq2000 platform. Each genome was processed by *Ngs-pipeline*, our bioinformatics tool developed for complete genome re-sequencing analysis, elimination of errors and prediction of the functional effects of single nucleotide and structural variations. In follow up comprehensive bioinformatics analysis of "3 Centenarian Genomes", we selected several genes bearing rare mutations in these centenarians and tested them in additional longevity population cohort of > 300 individuals (>85-90 years old), which included ~ 150 centenarians (from Russia and USA).

Results: We have identified about 3.5M of single nucleotide variations (SNV) in each genome and thousands of structural variations. For each SNV we checked it presence and minor allele frequency (MAF) in publically available datasets such as 1000 genomes project and Complete genomics. Among very rare (MAF < 0.1%) SNVs identified in each centenarian genome, several SNVs were notable in genes for specific biological pathway that highlight the importance of epigenetic regulation in longevity and elements of insulin/IGF-1 pathway linked previously to lifespan regulation in animal models.
Availability: Three centenarian genomes along with SNVs and structural variations will be available for browsing at our http://centenarian-browser.org/

Acknowledgements: This work was supported by Rostok Group.

MULTIPLE SOLUTIONS UNDER MODELING OF THE NITRATE UTILIZATION SYSTEM IN *ESCHERICHIA COLI*

Ri N.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: kashev@bionet.nsc.ru
** Corresponding author*

Key words: *nitrate reduction, nitrite toxicity, mathematical model, multistability*

Motivation and Aim: Nitrate is the most energetically favorable electron acceptor under anaerobic *E.coli* cell growth, reduction of that leads to formation of toxic nitrite. So regulation of genes involved in the nitrate-associated chain is closely related with nitrite metabolism genes, the most important elements of which are NrfA and NirB nitrite reductases. Regulation properties of the expression of *nrf* gene coding NrfA reductase [1] imitate substrate inhibition of the enzymatic reaction on the transcription level. It was theoretically shown that this type of regulation may results in bistability, i.e. to occurrence of two different stable states of the system [2]. To explore this possibility the model of nitrite utilization by *E.coli* cells during nitrate respiration was developed.

Methods and Algorithms: The Michaelis-Menten equations and generalized Hill functions were used to describe kinetic and molecular-genetic processes of the nitrate as well as nitrite metabolisms in *E.coli* cells [3]. Parameters of the model were taken from the published data or were estimated during model's adaptation to experimental data [1]. STEP+ was used for numerical calculations and analysis of the model [4].

Results: The model reproduces the conditions of *E.coli* cell culture in continuous chemostat during nitrate respiration and takes into account the properties genetic regulation of operons, coding the main nitrate, nitrite transporter (NarK) and enzymes catalyzed reduction of substrates: Nap, NRA, Nrf and NirB reductases [1]. Numerical analysis of the model dynamics in dependence of nitrate concentration in the chemostat has showed the existence of parameter set determining three stable and two unstable solutions caused by regulation properties of the Nap and Nrf expression. Switch from one stable solution to another is accompanied by significant change in intracellular nitrite concentration. The concentration value is exceeded the threshold of physiological stability of the strain used in [1]. This may cause spontaneous cell death.

Conclusion: It was firstly predicted that regulation properties of *E.coli* gene expression that involved in the nitrate-associated chain may result in the occurrence of different nitrate steady-state concentrations in the chemostat on basis of the mathematical modeling. The obtained results also predict the possibility of the spontaneous cell culture death under low rate of nitrite efflux for the strain used in [1] even under low nitrate steady-state concentration.

Acknowledgements: This study was partially supported by RFBR (№10-01-00717), Programs of the SB RAS Presidium (project №80), Programs of the Presidium RAS "Molecular and cell biology" (6.8) as well as "Biological diversity" (30.29) and Scientific school № 5278.2012.4

References:

1. Wang H., Gunsalus R.P. (2000) J. Bacteriol., 182(20): 5813-5822.
2. Chaudhury S., Igoshin O.A. (2009) J. Phys. Chem. B. 2009, V. 113. P. 13421-13428.
3. Likhoshvai V., Ratushny A. (2007) J. Bioinform. Comput. Biol., 5(2B):521-531.
4. Fadeev S. et. al. (2006) Proc. of the IC. on BGRS, Novosibirsk, 2:118-120

NEUROGENOMICS: CHALLENGES IN DEEP-GENOME STUDIES

Rogaev E.I.*[1,2,3,4], Reshetov D.A.[1,2], Tyazhelova T.V.[1] , Mikhaylichenko O.A.[1,2], Goltsov A.Y. [1,3], Gusev F.E.[1], Andreeva T.A.[1,3], Kaljina N.R.[1,3], Zenin A.A.[1], Protasova M.S.[1], Kunijeva S.[3], Grigorenko A.P. [1,3,4]

[1] Vavilov Institute of General Genetics, Moscow; [2] Lomonosov Moscow State University, Dep. of Bioengeneering and Bioinformatics; [3] Research Center of Mental Health, Russian Academy of Medical Sciences, Moscow, Russia; [4] Brudnick Neuropsychiatric Research Institute, Department of Psychiatry, University of Massachusetts Medical School, 01604, MA, USA

*e-mail: EVGENY.ROGAEV@umassmed.edu; * Presenting author*

Key words: *deep sequencing, SOLiD, schizophrenia, epigenomics, Alzheimer's disease*

Motivation and Aim: The identifications of genes or specific exogenous factors contributing to neuropsychiatric and behavioral diseases have been challenged. Interactions between environmental and genetic factors may underlie chronic neuropsychiatric disorders. Despite reduced reproductive fitness, the rate of incidence for schizophrenia and autism is relatively high in worldwide populations. We postulate that the individual genetic constitution attenuate the programmed epigenomic modifications during puberty. The study of regulatory sequences, including genes for non-coding RNA and the epigenomic regions marked by specifically modified chromatin, is a promising field in psychiatric genetics. To date the genome wide association studies (GWAS, employing relatively common SNPs across the genome) produced the genetic data with relatively low predictive and diagnostic values. The recently emerged concept that rare genetic variations, rather than common population variations, underlie common diseases challenges the standard genetic association approach in neuropsychiatric genetics. Direct sequencing of all genes, or preferably whole genome sequences, will provide most complete genetic information of the patient. However, from our current knowledge on population genetic variability, we expect millions of SNPs (single nucleotide polymorphisms) and up to thousand of CNVs per individual in comparison to reference genome sequence. Thus, excluding genetic "background" and identification of disease – related variations in the individual genomes require testing on experimental battlefield.

Methods and Results: We determined: (1) to our knowledge, the first complete genome sequences of patients with Alzheimer's disease and schizophrenia; 2) the complete genomes of familial cases of some neurological diseases with unknown mutant genes. (3) We also made the effort for development of experimental and bioinformatics methodology for identification of somatic mutations in human tissues. The data obtained by HiSeq2000, Solid and Pacific Bioscience platforms will be presented. The threshed criteria for filtering and identification of biologically significant private mutations and rare polymorphisms across the whole genome, selected chromosome or genetic locus linked to the disease have been developed. The role of non-coding RNAs, epigenetic–genetic interactions and genetic alterations, uniquely specific for *Homo sapiens* or occurring in both extinct and extant hominids, across epigenomic landscape will be discussed.

Acknowledgements: This work was supported by The Ministry of Education and Science of Russian Federation Federal target programs № 02.740.11.0854, 16.512.11.2083, 16.512.11.2102, FP7-HEALTH-2009, № 242257 (ADAMS), NIH/NIA 1R01AG029360, RFBR. We are grateful to Rostock group and A. Chikunov for support.

EVOLUTION OF LONG NON-CODING RNA GENES IN VERTEBRATES

Rogozin I.B.
IC&G SD RAS, Novosibirsk, Russia; NCBI/NLM/NIH, Bethesda MD, USA
e-mail: rogozin@bionet.nsc.ru

Mammalian genomes contain numerous genes for long non-coding RNAs (lncRNAs). The functions of the lncRNAs remain largely unknown but their evolution appears to be constrained by purifying selection, albeit relatively weakly. An obvious approach to gain insights into the mode of evolution and the functional range of lncRNA is to compare them with much better characterized protein-coding genes. For example, it is known that protein-coding genes that are under strong purifying selection. Analysis of human and mouse lncRNAs suggests that purifying selection acts on exons of lncRNA genes. The rate of evolution of protein-coding genes shows a universal negative correlation with expression: genes that are expressed highly and broadly across tissues typically are more conserved during evolution than genes with lower expression level and breadth. This universal negative correlation has been interpreted within the framework of the hypothesis of misfolding-driven protein evolution according to which misfolding is the principal cost incurred by protein expression. We sought to determine whether or not lncRNAs follow the same evolutionary trend and indeed detected a moderate but statistically significant negative correlation between the evolutionary rate and expression level of human and mouse lncRNA genes. Another property of vertebrate protein-coding genes is the almost perfect conservation of the exon/intron structure. Comparative studies of the exon/intron structure of lncRNA genes in various vertebrate genomes (including frog, chicken and fish) suggests that some lncRNA are conserved for over 500 million years.

T-CELL PROLIFERATION ON IMMUNOPATHOGENIC MECHANISM OF PSORIASIS: A CONTROL BASED THEORETICAL APPROACH

Priti Kumar Roy, Abhirup Datta
Center for Mathematical Biology and Ecology
Department of Mathematics
Jadavpur University
Kolkata -700032, India
e-mail: pritiju@gmail.com

Key words: *T-Cells, Dendritic Cells, Keratinocytes, Dermis, Epidermis, Cytokines, T-Cell Proliferation, Optimal Control*

Psoriasis vulgarism is a worldwide frequent autoimmune seditious skin illness differentiated by T-Cell agreeable hyperproliferation of Keratinocytes. The feature of T-Cell impeded continuous psoriatic lesions are the epidermal infiltration of essentially oligoclonal $CD8^+$ T-Cells and in all prospect also of $CD4^+$ T-Cells in the dermis. It was analyzed that, psoriatic scratches are piercingly distinguished, red and rather augmented lesions along with silver whitish scales. In this research article, we propose a mathematical demonstration for Psoriasis, involving a set of differential equations, relating to T-Lymphocyte Cells, Dendritic Cells and epidermal Keratinocytes. Here we incorporate the T-Cell proliferation in the system dynamics. Cell biological exploration on Psoriasis has been documented the suppression of epidermal T-Cell concentration. We are immensely paying attention to circulate, how the cell biological association modernizing through T-Cell proliferation in existence of control upshot taking position on T-Cell and Keratinocytes. We have originated a model corresponding to a miniature expression of Psoriasis together with available drug therapies. We also have considered that, T-Cells can be created by propagation of obtainable $CD4^+$ T-Cells. Through impulsive drug therapy, we have measured the effect of drug on the system dynamics. We have studied the model both in implicit and explicit approach. Our analytical and numerical outcomes reveal that, attributable to control effect, the Dendritic Cell population does not restrain in the margin between T-Cell and Keratinocytes, becoming stable after a certain time span. This illustration has been centered on the relations between T-Cells and Keratinocytes and impact of control upon them and would be able to provide a supremeperceptive of thedynamical systeminthepathogenesis ofPsoriasis.

References:

1. Roy Priti Kumar, Bhadra J., Comperative study of the suppression on T-cell and Dendritic cells in a mathematical model of Psoriasis, International Journal of Evolution Equation 5(2010), Issue-3, 309–326.

INTRON LENGTH DEPENDS ON PHASES OF SURROUNDING INTRONS

Roytberg M.A.*[1], Tsitovich I.I.[2], Astakhova T.V. [1]
Institute of Mathematical Problems of Biology, RAS, Moscow Region, Russia;
Institute of Information Transmission Problems, RAS, Moscow, Russia
e-mail: mroytberg@impb.psn.ru
** Corresponding author*

Key words: *exon, intron, intron phase*

Motivation and Aim: Phase of intron is a remainder of the total length of preceding exons divided by three. Type of intron is a triple xyz, where y is a phase of the intron under consideration, x is a phase of the preceding intron (0 for the first intron),, z is a phase of the subsequent intron (intron 0 for the last intron). The relation between lengths and phases of introns was studied in [1]. The aim of the work was to study relation between the length of the intron of its type.

Methods and Algorithms: We analyzed a set of insect and vertebrate introns, 17 organisms and 2036516 introns in total [ftp.ncbi.nih.gov/genomes]. The range of possible lengths of introns were divided into five intervals, the boundaries of the intervals were different for vertebrates and insects, because of difference in the avearage lengths of the introns. Given sample of introns and partition of all possible intron lengths, an R-value is a ratio of number of introns with lengths from the maximal length interval to the number of introns with lengths from the minimal length interval. For each organism examined the R-value was calculated for two samples of introns: (1) all introns (2) start introns

Results: For each species the R-values for different types were ordered by decrement, the results are given in the table. For all mammals and birds the best values are achieved for same intron types; the "R-value" column shows average values for mammals ad birds resp.

	First MAX		Second MAX		Third MAX		Fourth MAX	
Organizms	Type	R-value	Type	R-value	Type	R-value	Type	R-value
Mammals	0 1 1	34.41%	0 2 1	31.29%	0 0 1	29.56%	2 1 1	25.03%
Birds	0 1 1	17.46%	0 2 1	13.45%	0 0 1	13.03%	2 1 1	11.26%
Xenopus	0 1 1	21.91%	0 2 1	15.07%	0 0 1	14.17%	2 1 1	11.41%
Danio rerio	0 1 1	14.91%	2 1 1	11.69%	0 2 1	11.68%	0 0 1	9.85%
Lizard	0 1 1	52.45%	0 2 1	42.17%	0 0 1	37.57%	2 1 1	30.56%
Insects	0 1 1	12.48%	0 0 1	8.26%	0 2 1	7.74%	1 1 1	7.57%

For the start introns in all species except mammals and xenopus the maxomal R-value corresponds to the type 0 1 1; for mammals and xenopus it corresponds to 0 2 1.

Conclusion: Average intron length depends not only its phase but also of phases of intron neighbors. The maxima of R-values was observed for the introns of type X1Y and XY1.

Availability: http://lpm.org.ru/~mroytberg/intron_phase

References:

1. T. Astakhova, I. Tsitovich, M. Roytberg. Proceedings of the International Moscow conference on computational molecular biology MCCMB'11 Moscow, Russia, July 21-24, 2011. c.321-322.

TARDIVE DISKINESIA AND POLYMORPHISM OF PHOSPHATIDYLINOSITOL- 4-PHOSPHATE 5-KINASE IIA GENE IN RUSSIAN SCHIZOPHRENIC PATIENTS

Rudikov E.V. *[1], Gavrilova V.A. [1], Fedorenko O.Y. [1], Boyarko E.G.[1], Semke A.V. [1], Sorokina V.A. [2], Govorin N.V. [3], Ivanova S.A. [1]

[1] *Mental Health Research Institute SB RAMSci, biological psychiatry, Tomsk, Russia;*

[2] *Kemerovo Regional Clinical Psychiatric Hospital, psychiatry, Kemerovo, Russia;*

[3] *Chita State Medical University, psychiatry, Chita, Russia*

e-mail: korvin_w@mail.ru

** Corresponding author*

Background: Pharmacogenetic studies of tardive dyskinesia are very promising direction to develop individualized antipsychotic treatment. Phosphatidylinositol-4-phosphate-5-kinase IIa (PIP5K2A) is one of the key enzyme in phosphatidylinositol-4,5-bisphosphate biosynthesis, plays an important role in membrane transduction of neurotransmitter signals and in intracellular signaling. PIP5K2A gene is located in schizophrenia candidate region on chromosome 10p14-11. Polymorphisms of this gene have been shown to be associated with schizophrenia in European and Chinese populations, but there were no such studies in Russia.

Objective: Our study aimed to investigate association of (N251S)-PIP5K2A (rs10828317) polymorphism with tardive diskinesia in Russian schizophrenic patients.

Materials and methods: Blood samples from 355 Russian Caucasian patients with clinically established schizophrenia (with an age of 43±15.8 years) were taken in four different psychiatric departments in West Siberia. Abnormal Involuntary Movement Scale (AIMS) was used to assess tardive dyskinesia cross-sectionally. Control group consisted of 100 healthy volunteers. Genotyping of (N251S)-PIP5K2A (rs10828317) was performed on ABI StepOne Plus with TaqMan1 Validated SNP Genotyping Assay (Applied Biosystems). The program SPSS11.5 was used for statistical analysis. Hardy–Weinberg equilibrium (HWE) and differences in genotype frequencies were tested using a chi-square test. Comparisons of AIMS-score in different groups were carried out with Kruskal Wallis test and Mann-Whitney test with Bonferroni correction.

Results: The genotype distribution of (N251S)-PIP5K2A (rs10828317) polymorphism was in agreement with HWE ($\chi 2$ =0.32, p=0.6481) in control group, but there was a disequilibrium in group of schizophrenic patients ($\chi 2$ =9.06, p=0.0028). 40% of patients and 45% of healthy volunteers were homozygous for the T-allele, 40.3% of patients and 46% of control persons were heterozygous, and 19.7% of patients and 9% of healthy volunteers were homozygous for the C-allele ($\chi 2$ =6.25, p=0.044). We found an association of rs10828317 with schizophrenia (p=0.04, Odds Ratio=2.48, 95%CI=1.19-5.17 for the CC genotype). CC-carriers with schizophrenia also had a higher mean AIMS score (6(2-9)) (median (25%-75% percentiles)) in comparison to those with the CT (3(0-6)) or the TT (2 (0-4)) genotype (Kruskal Wallis test – 24.74, p<0.0001, Mann-Whitney test with Bonferroni correction - p(CC/TT)<0.0001, p(CC/CT)=0.0009, p(TT/CT)=0.09). Subsequently, frequency of CC-carriers was significantly higher in group of schizophrenic patients with tardive diskinesia compared with the group of schizophrenic patients without tardive diskinesia (34.4% and 15.7% respectively, $\chi 2$ =15.4, p=0.0004, and OR=2.81 95%CI=1.61-4.91 for the CC genotype, p=0.0005).

Discussion: (N251S)-PIP5K2A (rs10828317) is known to be a functional mutation. Previous studies show that mutant kinase inefficient to activate the KCNQ channels that may lead to lack of dopaminergic control in schizophrenic patients [1]. Moreover, (N251S)PIP5K2A decreased membrane abundance of excitatory amino acid transporter EAAT3 in study on EAAT3-expressing oocytes and human embryonic kidney cells [2]. Taken together, these facts may act as biological explanation of association of (N251S)-PIP5K2A (rs10828317) with tardive diskinesia and schizophrenia.

Conclusions: The significant association of (N251S)-PIP5K2A polymorphism with tardive diskinesia has been found. CC-carriers with schizophrenia had higher risk of tardive diskinesia and more severe symptoms as evaluated by AIMS. Further studies are needed to support our findings.

CIRCULATING DNA IN CANCER PATIENTS BLOOD: GENERAL CHARACTERISTICS AND WHOLE – GENOME ANALYSIS

Rykova E.Y.*[1], Morozkin E.S.[1], Loseva E.M.[1], Skvortsova K.N.[1], Ponomaryova A.A.[2], Kurilshikov A.M., Morozov I.V., Bryzgunova O.E.[1], Bondar A.A.[1], Zaporozhchenko I.S.[1], Kapitskaya K.Y.[3], Azhikina T.L.[3], Cherdyntseva N.V.[2], Vlassov V.V.[1], Laktionov P.P. [1]

[1] *Institute of Chemical Biology and Fundamental Medicine, SB RAS, Novosibirsk, Russia;*

[2] *Cancer Research Institute, SB RAMS, Tomsk, Russia;*

[3] *Shemyakin and Ovchinnikov Institute of Bioorganic Chemistry, RAS, Moscow, Russia*

e-mail: rykova@niboch.nsc.ru

** Corresponding author*

Key words: *circulating DNA, methylation, cancer, microarray, next generation sequencing*

Motivation and Aim: DNA modification studies in circulating DNA (cirDNA) may lead to development of specific non-invasive cancer biomarkers. CirDNA pool complexity became evident indicating that discovery of extracellular DNA generation and circulation patterns in cancer is essential to develop the valid diagnostic markers.

Methods and Algorithms: Comparative study of plasma cirDNA methylation modifications from prostate cancer patients and two control groups was made using the epigenome-wide screening by hybridization to the Human CpG island microarrays (UHN, Toronto, ON). Methylation index of RARB2 gene changes in cirDNA from lung cancer patient blood were detected using methylated DNA fragment enrichment by Methylated CpG Island Recovery Assay (MIRA) followed by real-time PCR. Apoptotic DNA isolated from culture medium (cm-apoDNA) of human umbilical vein endothelial cells (HUVEC) induced to apoptosis was compared with genomic DNA (gDNA) from the same normal cells using SOLiD 3 platform (Applied Biosystems, USA).

Results: Using the Differential Methylation Hybridization method 39 prostate cancer-associated changes in cirDNA methylation were identified. Pyrosequencing analysis of 7 selected loci revealed aberrant CpG methylation of two novel candidate cancer markers (ZC3H4 and RNF219). Gene RNF219 methylation was evaluated by cloning and sequencing of individual cirDNA molecules which demonstrated its diagnostic potential. A modified techniques for the methylated CpG-containing DNA fragment enrichment was developed based on the methyl-CpG affinity binding with the recombinant protein containing human methyl-binding domain 2 (MBD2) fused with glutathione S-transferase. Using the developed techniques RARB2 gene methylation index changes were found in cirDNA from lung cancer patients. As far as apoptosis is the main source of cirDNA, apoptotic DNA from HUVEC cells induced to apoptosis and genomic DNA from untreated cells were compared using SOLiD 3 platform. The representation analysis of repetitive elements revealed that cm-apoDNA is significantly enriched with Alu-repeats and depleted with LINE-1 elements compared with genomic DNA from intact cells. These data and the location of Alu-repeats mainly in euchromatin regions demonstrate the enrichment of cm-apoDNA with transcriptionally active gene-rich DNA sequences.

Conclusion: CirDNA genomics and epigenomics study provide promising source for development of non-invasive cancer biomarkers.

Acknowledgements: The research has been carried out with support of the grants from RFBR № 11-04-12105-offi-m-2011, SB RAS Program in collaboration with other scientific organizations № 65, RAS Program “Fundamental Science for Medicine” № 23.

CONSTRUCTION AND ANALYSIS OF THE PROTEIN-PROTEIN INTERACTION NETWORK FOR THE SPERMATOZOA

Sabetian S.F.J.*[1], Lau C.[1], Bostan H.[1], Valipour A.R.[2], Shamsir M.S.[1]
Faculty of Bioscience & Bioengineering[1], Faculty of Civil Engineering[2], Universiti Teknologi Malaysia, Malaysia
e-mail: sudsabet@yahoo.com
** Corresponding author*

Key words: *protein-protein interaction map, spermatozoa, sperm-egg membrane fusion, male infertility*

Motivation and Aim: Currently, protein-protein interaction network is not available for the human spermatozoa. The main purpose of this study is to map all the protein that is associated with the spermatozoa. The creation of this map will allow a more detailed examination of functional clusters and elucidate any relationships between the protein nodes. We believe that the creation of the interaction map would contribute to the understanding of sperm's proteomic functional role in sperm-egg membrane fusion (1), male-factor infertility, sperm dysfunctions (2) and drug targets for infertility treatment (3).

Methods and Algorithms: Cytoscape 2.8.2 was used to construct and analyze the protein-protein interaction (PPI) network for all proteins in the sperm's function. We have classified all of the proteins of this network by using Allegro-Mcode algorithm and predicted their possible biological processes by evaluating the results of GO enrichment analysis, using the Cytoscape BinGo plugin.

Results: The protein network consists of 6452 protein nodes and 36319 interactions among those proteins. From the 6452 input proteins of this map, only 1065 proteins as 62 clusters were assembled. The G-protein coupled receptor activity, transmembrane receptor activity and peptide receptor activity: G-protein coupled functions are the most significant nodes in molecular function map of important cluster. The ranking of clusters of the spermatozoa PPI network highlighted the importance of LIS1, CLIP1, PLK1, and CENPH in sperm-egg membrane fusion as well as 5-HT-1B, CHRM2, DRD2 and DRD3 in drug binding process.

Conclusion: We have built the first protein-protein interaction network of the spermatozoa using the computational approach. The analysis of the protein network enabled us to identify a set of important proteins in the PPI network. Although these candidates will still require laboratory validation, the establishment of these protein sets will allow a smaller focus on the essential protein and their functional roles.

Availability: PPI for spermatozoa is available at http://birg4.fbb.utm.my/spermatozoamap.

References:

1. K.Kaji and A.Kudo. (2004) Focus on Fertilization: The mechanism of sperm–oocyte fusion in mammals, *Reproduction,* **127:** 423-429.
2. I.A.Brewis *et al.* (2005) The spermatozoan at fertilization: Current understanding and future reseach directions. *Progress in Drug Research,* **8(4):** 241-251.
3. M.Strong and D.Eisenberg. (2007) The protein network as a tool for finding novel drug targets. *Progress in Drug Research,* **64:** 193-215.

INTRIGUING STRUCTURES IN TRIPLET DISTRIBUTION ALONGSIDE A GENOME

Sadovsky M.G.*, Mirkes E.M.
Institute of Computational Modelling of SB RAS, Krasnoyarsk, Russia
e-mail: msad@icm.krasn.ru;
** Corresponding author*

Key words: *order, periodicity, correlation, function, sense*

Motivation and Aim: The aim of the study is to identify, decipher and explore the semantics of a new structure revealed from the nucleotide sequences. The structure is expected to have a deep connection to the functionality and semantics of the nucleotide sequences. Simultaneously, similar relation is expected to be revealed between this newly revealed structure, and taxonomy of the bearers of those nucleotide sequences.

Methods and Algorithms: The distribution of the triplets (or, wide, short strings) alongside a sequence was developed. The distribution was defined as a distance to the nearest neighbour, i.e., for two triplets ω_1 and ω_2 the number n_l of nucleotides between them was counted so that there is no other word ω_2 embedded somewhere inside the string of the length n_l. The longest distance to detect the nearest neighbour was as long, as 10^4 nucleotides. The distribution function was developed for all 4096 couples of triplets. Two kind of surrogate sequences have been developed to distinguish the biological effects from the combinatorial ones: the former was random non-correlated sequence with the same composition of nucleotides, and the latter was a Markov chain sequence of order 3 to 6.

Results: A number of chromosomes of human genome, and chimpanzee genome have been studied. All chromosomes exhibit a strong and extremely unusual structures in the distribution of the triplets (to the nearest neighbour). Thus, an explicit and strong periodicity in CCC – GGG triplets has been found, with the period of 13 and 36 nucleotides. Some other couples exhibit more complex and long-range correlations (up to 250 nucleotides). The pattern of the couples distribution strongly differs from similar observed for random non-correlated sequence, and from those observed over Markov chain surrogate sequence of the order varied from 2 to 6.

Conclusion: a new structure is found in nucleotide sequences. Very complicated patter of the couples distribution is peculiar for higher eukaryotic organisms, while unicellular eukaryotes, as well as bacteria exhibit very smooth pattern close to that one observed for Markov chains of relevant order.

OLIGONUCLEOTIDE FREQUENCIES AND GC CONTENT OF BACTERIAL GENOMES ARE RELATED TO THE ENVIRONMENT EVOLUTION

Safronova N.S.[1], Suslov V.V.[2], Afonnikov D.A.[1, 2], Podkolodnyy N.L.[2, 3], Mitra C.K. [4], Orlov Y.L.*[1, 2]

[1] *Novosibirsk State University, Novosibirsk, Russia;*
[2] *Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia;*
[3] *Institute of Computational Mathematics and Mathematical Geophysics, SB RAS, Novosibirsk, Russia;*
[4] *University of Hyderabad, India*
e-mail: orlov@bionet.nsc.ru
** Corresponding author*

Key words: *evolution, genomics, oligonucleotide distribution, GC content*

Motivation and Aim: GC content and genome size are one of several integral genome features limiting potential spectrum of licenses (habitats) of prokaryotic taxons [1]. GC content of bacterial genomes varying between 25% and 75%, has been investigated long time and many hypotheses have been put forward to explain GC content variation and its relationship to other fundamental processes. Codon usage bias, especially GC content at the third codon position, correlates with the trend of GC content variations, and may be driven by GC content changes [2]. Another important observation is unequal distribution of oligonucleotide frequencies in coding regions, regulatory regions and complete genomes.

Methods and Algorithms: We used annotated and assembled compete genome sequences of prokaryotic organisms downloaded from NCB ftp-site. We have used also transcription factor binding sites from JASPAR database. Using in-house computer programs we counted oliconucleotide frequencies, GC content and correlations [3].

Results: We revealed their associations *in silico* and show taxa biasing common trend. Relation to ecology and early prokaryotic evolution will be discussed. We have selected common oligonucleotides for bacterial species. It is interesting to observe skewed distribution of oligonucleotide frequencies for all the genomes studied and robustness of power law disctribution.

Conclusion: Environmental or bacteriological factors, such as genome size, temperature, oxygen requirement, and habitat, either play subsidiary roles or rely indirectly on different factors to fine-tune the GC content. These results provide an insight into mechanisms of GC content variation in adapting to changing environments.

Acknowledgements: The work is supported in part by RFBR 11-04-01888, 11-04-92712-IND, 12-04-92702-IND, the Russian Ministry of science and education (projects No. 07.514.11.4023, 857), IP SB RAS 21, 39, 130, Presidium RAS Programs No. 6.8, 28.

References:

1. V.V. Suslov et al. (2012) Genome features and GC content of prokaryotic genomes are related to the environment evolution, *Paleontology Journal (Mosk),* (In press).
2. H. Wu et al. (2012) On the molecular mechanism of GC content variation among eubacterial genomes, *Biol Direct,* 7**:** 2.
3. 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.

IDENTIFICATION OF NEW DERIVATIVES OF OKADAIC ACID - SELECTIVE INHIBITOR OF PROTEIN PHOSPHATASE 1 (PP1) AND 2A (PP2A)

Samofalova D.A.*, Karpov P.A., Blume Ya.B.
Institute of Food Biotechnology and Genomics, Natl. Academy of Sci. of Ukraine, Kyiv
e-mail: samofalova.dariya@gmail.com
** Corresponding author*

Key words: *protein phosphatases, okadaic acid, derivatives, chemoinformatics*

Motivation and Aim. Okadaic acid (OA) strongly inhibits protein serine/threonine phosphatases PP1 and PP2A/2B [1]. In comparison with the other inhibitors the inhibitory effect of OA is strongest for PP2A, followed by PP1, and then PP2B [2]. This toxin is now used as powerful research tool for an increasingly wide variety of cellular events regulated by reversible protein phosphorylation [3]. Because of the lack of highly PP1 and PP2A/2B selective inhibitors, design and search of new biologically active OA derivatives is extremely important.

Methods and Algorithms. Modeling of PP1 and PP2A spatial structure was performed using SWISS-MODEL service (http://swissmodel.expasy.org/). The *ligands for docking* input were prepared using CCDC Hermes. Flexible docking of OA derivatives was *performed* using CCDC GOLD Suite 5.1. Binding site for PP1 was specified in 15 E radius about -NE2 (HIS125), and for PP2A in 20 E radius about -ND2 (ASN117). For docking evaluation CCDC GOLD scoring functions were used (ChemScore, GoldScore and ASP) and results of molecular dynamics simulations in GROMACS.

Results and conclusion. Based on template structures of human PP1 (PDB: 1U32) and PP2A (PDB: 2IE4) in complex with OA we reconstructed 3-D models of homologous proteins from *Arabidopsis thaliana*, *Emericella nidulans* and *Salmonella typhimurium*.

A high sequence and structure identity of protein phosphatases of different origin allow us to conclude similarity of OA binding sites in molecules of PP1 and PP2A. Based on chemoinformatic analysis of PubChem (http://pubchem.ncbi.nlm.nih.gov) and ZINK (http://zinc.docking.org/) databases, 26 derivatives of OK were selected. Complexes with selected OA derivatives were predicted by flexible docking and evaluated based on CCDC GOLD scoring functions and results of molecular dynamics in GROMACS. As a result, we selected five compounds not previously described as potential PP1 and PP2A inhibitors.

Acknowledgments: This work was supported by STCU#5215 grant: “Search of effective protein phosphatases inhibitors using nanochemical approaches and evaluation of their biological activity *in silico*”.

References:

1. A. Garcia, X. Cayla, J. Guergnon, et al. (2003) Biochimie, 85 (8): 721–726.
2. R.E. Honkanen (1993) FEBS Letters, 330 (3): 283–286.
3. C.F.B. Holmes, M.P. Boland (1993) Cur. Opin. Struct. Biol., 3: 934–943.

A MAP OF ANAPHASE CHROMOSOMAL BREAKS INDUCED BY CONDENSIN LOSS

Samoshkin A.[1], Dulev S.[2], Loukinov D.[3], Rosenfeld J.A.[4], Strunnikov A.V. *
** Guangzhou Institutes of Biomedicine and Health, Molecular Epigenetics Laboratory, 190 Kai Yuan Avenue, Science Park, Guangzhou 510530, Guandong, China*
[1] NIH NCI, Genome Structure and Function Section, Bethesda, MD, USA;
[2] Ludwig-Maximilians-Universitծt, Adolf Butenandt Institut, 80539 Munich, Germany;
[3] NIH NIAID, Laboratory of Immunopathology, Rockville, MD, USA;
[4] Division of High Performance and Research Computing, University of Medicine & Dentistry of New Jersey, Newark, NJ, USA

Abstract: Condensin complexes are essential for mitotic chromosome condensation and segregation, while condensin dysfunction leads to chromosomal bridging in mitosis, paving the way for rapid genome destabilization undetectable by checkpoints, similar to genome rearrangements found in many cancer genomes. To map potential double-strand breaks specifically occurring in late anaphase, human chromosomes depleted of condensin were analyzed by gamma-H2AX ChIP-seq. Condensin-depleted chromosomes from HeLa cells contained distinct gamma-H2AX enrichment zones, 75% of which overlapped with known hemizygous deletions in cancers. Furthermore, some tandemly repeated DNA sequences, analyzed by ChIP-seq and custom repeat array ChIP-chip using independent high-throughput and bioinformatic approaches, showed significant gamma-H2AX enrichment in condensin-depleted anaphases. The preferential targets of such an enrichment included simple repeats, centromeric satellites, and rDNA. The genomic regions that are specifically destabilized upon condensin dysfunction may constitute a quantifiable condensin-specific CDP (Chromosome Destabilization Pattern).

MATHEMATICAL MODEL OF AUXIN RESPONSIVE REPORTER DR5 ACTIVITY IN PLANT CELL

Savina M.S. [1, 2], Mironova V.V.*[1], Akberdin I.R. [1], Omelyanchuk N.A. [1], Likhoshvai V.A.[1, 2]
[1] *Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia*
[2] *Novosibirsk State University, Novosibirsk, Russia*
e-mail: antonec@yandex.ru
** Corresponding author: kviki@bionet.nsc.ru*

Key words: *auxin response, DR5, mathematical model*

Motivation and aims: Hormone auxin is the main regulator of plant growth and development. The DR5 reporter lines (DR5::GFP, DR5::GUS and others) are used in many researches to explore auxin distribution in plant tissues. The DR5 synthetic promoter contains 7 repeats of auxin responsive sites AuxRE (with the sequence cctttTGTCTC). AuxRe are the binding sites for transcription factors of ARF family, which mediate the primary auxin response. Despite being widely used, the mechanism of auxin responsive transcription from DR5 promoter is not well understood. Here we used the method of mathematical modeling for investigation of possible regulatory mechanisms of DR5 activity.

Methods: Mathematical models of DR5 activity have been created in terms of generalized Hill functions [1]. Numerical calculations were performed in MGSModeller [2] and Mathematica packages. For parameter estimation we used gradient descent method. As a function of minimization we used euclidean distance of theoretical calculations to experimental data published in [3]. Minimization was performed in Mathematica package. Numerical analysis of combinatory variants of DR5 models was performed in MGSModeller [2].

Results: We created a number of mathematical models of auxin-dependent regulation of DR5 expression, which includes different types of competitive, uncompetitive as well as composite interactions of the regulatory AuxRE sites in auxin response. The models were tested on the experimental data [3], where a relative activity of $DR5_N$ promoters was measured and revealed depending on the number of AuxRE sites in the promoter (N=1,…,8). Among more than hundred variants, the one model was selected giving the best qualitative agreement of the numerical calculation to the experimental data [3].

Conclusion: By numerical analysis we propose the following model of auxin responsive DR5 activity: (1) auxin influence DR5 activity via complex cooperatively-competitive mechanism; (2) DR5 promoter has low but not zero basal (auxin-independent) activity; (2) auxin-concentration effect realized through one AuxRE inhibits DR5 transcription; (3) auxin-concentration effect realized through N-2 or N AuxREs activates DR5 transcription.

Acknowledgements: Numerical calculations were performed on supercomputer cluster of Shared Facility Center Bioinformatics SB RAS. The work was supported by the RFBR grants № 11-04-01254a and №10-01-00717, Integration projects SB RAS № 80, The Russian President grant SS-5278.2012.4 and RAS programs B.27, B.25.

References:

1. Likhoshvai V, Ratushny A. (2007). *Journal of Bioinformatics and Computer Biology 5: 521–531.*
2. Kazantsev F.V. et al. In this proceedings.
3. Ulmasov T. *et al*. (1997). *The Plant Cell*, **9**: 1963-1971.

ON THE FACTOR ANALYSIS OF MASS CELL MOVEMENTS IN AMPHIBIAN GASTRULATION

Scobeyeva V.A., Cherdantsev V.G.

Motivatioan and aim: Factor analysis generally exploits only for separating of the potentially autonomous components of a process with no respect to their physical or even biological nature. Our idea was to use factor analysis as a tool for reconstructing mass cell movement trends in amphibian gastrulation. We assumed that factor loadings of the quantitative morphological variables entering the same factor separated by the factor analysis were components of the same mass cell movement vector.

Methods and algorithms: We analyzed the normal variability of gastrulation in two related amphibian species in both the embryos fixed at successive stages of gastrulation and living embryos with the aid of time-lapse recording of the individual developmental pathways in three Anuran species – *R. temporaria*, *R. arvalis* and *Pelobates fuscus*.

Results: We succeeded to separate principally the same components of the process that embryologists have separated by the traditional methods and essentially specify their spatiotemporal dynamics and mechanical forces operating on these processes. The amphibian gastrulation is a series of cycles embedded one into another: gastrulation begins with and ends by the epiboly, while embedded into this cycle are the planar convergence of cell flows and lateral dorsal lip spreading being of a crucial importance for the formation of the embryonic axial structures. Almost irrespectively to the number of variables under consideration, the overall variance distributes between no more than two factors, the first factor corresponding to a movement vector dominating at a given gastrulation stage. The second factor corresponds to the initiation of a new system of morphogenetic correlations that will dominate at the next gastrulation stage, or to the remnants of a system that has dominated at the earlier stage of gastrulation

Conclusion: As at follows from the factor analysis, the formation of a new system of correlations passes through a standard sequence of stages. At a first stage, the components of future correlation systems vary independently, then they integrate to become components of the same mass cell movement and then the system breaks again to independently varying components. Thus, lying behind the linear succession of developmental stages are the repetitive transformations of the vector field of mass cell movements.

PROFILE OF THE CIRCULATING RNA IN APPARENTLY HEALTHY INDIVIDUALS AND NON-SMALL CELL LUNG CANCER PATIENTS OBTAINED WITH MASSIVELY PARALLEL SEQUENCING OF TOTAL BLOOD PLASMA RNA

Semenov D.V.*[1], Baryakin D.N.[1], Brenner E.V.[1], Kurilshikov A.M.[1], Kozlov V.V.[2], Narov Y.E.[2], Vasiliev G.V.[3], Bryzgalov L.O.[3], Chikova E.D.[1], Filippova J.A.[1], Kuligina E.V. [1], Richter V.A.[1]

[1] Institute of Chemical Biology and Fundamental Medicine SB RAS, Novosibirsk, Russia;
[2] Novosibirsk Regional Cancer Centre, Novosibirsk, Russia;
[3] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
** Corresponding author*
e-mail: semenov@niboch.nsc.ru

Key words: c*irculating RNA, massively parallel sequencing, human blood plasma RNA, non-small-cell lung cancer*

Motivation and Aim: The understanding of circulating RNA structures and functions expands fundamental knowledge of cell communications and signaling pathways as well as allows developing new molecular diagnostic approaches. The aim of this study was to document profile of common and peculiar RNA species normally circulating in blood of healthy individuals and of patients with non-small cell lung cancer with massively parallel sequencing of human blood plasma RNA.

Methods and Algorithms: Total RNA was extracted from blood plasma samples of 8 apparently healthy individuals and 8 patients with non-small cell lung cancer. To obtain comprehensive cDNA libraries RNA was dephosphorylated and then 5'-phosphorylated. 5'-phosphorylated total plasma RNA was ligated with adapters, reverse transcribed and 16 personalized cDNA libraries were constructed. Libraries were amplified and sequenced with SOLiD™ system. The sequenced 35-nt-long reads were mapped to human transcriptome/genome, classified and quantified with Bowtie/Cufflinks software.

Results: Fragments of rRNA, mitochondrial transcripts, microRNAs, fragments of scRNAs, snRNAs and snoRNAs, fragments of several mRNAs as well as the set of newly discovered transcripts were found to be permanent representatives of human blood plasma RNAs. Comparison of circulating RNA profiles of healthy subjects and cancer patients allowed us to document diagnostically significant RNA species, including fragments of mRNA, miRNA and others non-coding regulatory RNAs.

Conclusion: Documented profile of circulating RNA of healthy individuals and patients with lung cancer provides the basis for development of new research of circulating regulatory RNAs and allows to construct new platforms for early diagnosis of human malignancy.

Acknowledgements This study was supported by Integration SB RAS Grant #18, RFBR grants #10-04-01386-a and #10-04-01442-a.

COMPUTATIONAL AND ANALYTICAL ASPECTS OF A NEW COMPLEX MODEL DESCRIBING HUMAN CARDIOVASCULAR SYSTEM

Semisalov B.V.*[1, 2], Kiselev I.N. [1, 2], Sharipov R.N. [2, 3], Kolpakov F.A. [1, 2]
[1] *Design Technological Institute of Digital Techniques SB RAS, Novosibirsk, Russia;*
[2] *Institute of Systems Biology, Ltd, Novosibirsk, Russia;*
[3] *Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia*
e-mail: vibis@systemsbiology.ru, vibis@ngs.ru
** Corresponding author*

Key words: *human cardiovascular system, complex model, numerical results*

Motivation and Aim: Essential hypertension is on the first place of death rate in the developed countries. That is why a number of models of human cardiovascular system (CVS) has been proposed for searching the possibilities for optimisation of the disease treatment. In this work we consider 1D hemodynamic model (HDM) of the human arterial tree [1] and propose approach for combining it with short-term and long-term ordinary differential equation models of the human CVS [2,3].

Methods and Algorithms: The HDM represents a system of partial differential equations obtained from Navier-Stokes equations. For searching the numerical solutions of this model we used the specific tree packing algorithm and orthogonal sweep method developed in Sobolev's Institute of Mathematics [4].

Results: All the mentioned models were implemented using the BioUML platform (www.biouml.org). An essential problem was to specify physiologically adequate and mathematically correct boundary conditions on the loose ends of the HDM tree. After a careful study we have found an effective way for determining these conditions using the models from [2,3]. This way utilizes simple physical laws like Poiseuille's, filtration and mass conservation laws and additional assumptions allowing to combine three models in one complex model of CVS. The complex model created in the BioUML platform is simulated on the principles of agent based modeling. It can reproduce different pathologies and can be used for studying the causes of different CVS diseases.

Conclusion: We have developed initial version of the complex CVS model. Although the validation of some parameters and principles remains to be done in cooperation with data obtained from medicine, the main part of the model construction is complete.

Availability: www.biouml.org/cvs.shtml

Acknowledgements: This work was supported by the.project № 91 of SB RAS.

References:

1. D. Lamponi (2004) One dimensional and multiscale models for blood flow circulation. Pour l'obtention du grade de docteur es sciences. Ecole Polytechnique Federale De Lausanne.
2. A. P. Proshin and Yu.V. Solodyannikov (2006) Mathematical modeling of blood circulation system and its practical application, Automat Remote Contr, 2: 174–188.
3. F. Karaaslan et al. (2005) Long-term mathematical model involving renal sympathetic nerve activity, arterial pressure, and sodium excretion, Ann Biomed Eng, 33(11): 1607-1630.
4. Cardiovascular system and arterial hypertension: biological, genetic and physiological mechanisms, mathematical and computer modeling, L.N. Ivanova (Ed.) (2008). Novosibirsk, Publishing house of the SB RAS, 252 pp.

BioUML: PLUGIN FOR STOCHASTIC MODELING OF BIOLOGICAL SYSTEMS

Semisalov B.V.*[1, 2], Kiselev I.N. [1, 2], Sharipov R.N. [2, 3], Kolpakov F.A. [1, 2]
[1] *Design Technological Institute of Digital Techniques SB RAS, Novosibirsk, Russia;*
[2] *Institute of Systems Biology, Ltd, Novosibirsk, Russia;*
[3] *Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia*
e-mail: vibis@systemsbiology.ru, vibis@ngs.ru
** Corresponding author*

Key words: *stochastic approach; exact, approximate, hybrid algorithms*

Motivation and Aim: It is known that the deterministic approaches for modeling of biological systems fail in some cases. For example, the dynamics of virus spreading or chemical reaction with small number of interacting molecules cannot be described with ODEs. The stochastic nature of those processes does not allow to determine whether the epidemic will start or not, and the ordinary assumptions of chemical kinetics cannot work when the number of molecules of reactants is low. That's why the stochastic modeling comes to the fore.

Methods and Algorithms: The stochastic simulation algorithms are divided into three groups: exact, approximate and hybrid algorithms. The exact algorithms based on Gillespie and Gibson-Bruck methods fires only one reaction on each time step [1], giving the most precise description of stochastic nature of process. But if the number of interacting becomes high, the run-time of exact methods becomes too large. The approximate modified tau-leaping method [2] allows to overcome this problem by firing the reaction multiple times at each step. On the other hand, if we have two types of reactants with small and large number of molecules the mentioned methods cannot be efficient. In this case we have to use the hybrid methods like the "maximal time step method" [3] separating out the fast and slow processes.

Results: The algorithms mentioned above were implemented as a plug-in for the BioUML platform – the open source integrated Java platform for visual modeling and analysis of complex biological systems. These algorithms were modified and improved to observe the results of simulation in arbitrary time point and support the discrete events in a model. All the algorithms have passed successfully the SBML discrete stochastic models test suite (see www.dsmts.google.com).

Conclusions: The functional possibilities of BioUML in modeling have been extended by the efficient tool for stochastic modeling of biological systems. Our next goal is to implement a hybrid algorithm uniting the stochastic and deterministic approaches.

Availability: The stochastic solvers are available as a part of the BioUML platform at www.biouml.org.

References:

1. M. A. Gibson, J. Bruck (2000) Efficient Exact Stochastic Simulation of Chemical Systems with Many Species and Many Channels, *J. Phys. Chem. A*, 104 (9): 1876-1889.
2. C. Yang et al. (2006) Efficient Stepsize Selection for the Tau-Leaping Simulation Method, *J. Chem Phys,* 124(4): 044109.
3. J. Puchalka, A. M. Kierzek. (2004) Bridging the Gap between Stochastic and Deterministic Regimes in the Kinetic Simulations of the Biochemical Reaction Networks, *Biophys J,* 86(3):1357-1372.

TOWARDS AN ANALYSIS OF THE STRUCTURE OF THE SHORT ARM OF 5B CHROMOSOME OF THE BREAD WHEAT *TRITICUM AESTIVUM* L.

Sergeeva E.M.*, Afonnikov D.A., Bildanova L.L., Koltunova M.K., Timonova E.M., Salina E.A.
Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
e-mail: sergeeva@bionet.nsc.ru
** Corresponding author*

Motivation and Aim: *Triticum aestivum* (bread wheat) is the allohexaploid (genomic formula BBAADD, 2n = 6x = 42), with the large genome size of 17 000 Mbp. Approximately 90 % of wheat genome account for repetitive sequences of different origin and degree of reiteration that make the wheat genomic sequencing very difficult and expensive. At the present day the wheat genome sequencing is in progress and little is known concerning the structure of the short arm of chromosome 5B. The basic approach for wheat genomic sequencing is the sequencing of minimal set of overlapping clones from chromosome-specific BAC-libraries. The key stage is the anchoring of mapped molecular markers to the BAC-clones. The essential task is to obtain the markers distributed uniformly along all the chromosome length. The 454-sequencing technology (Roche) produces the reads long enough for development such kind of PCR markers named ISBP (Insertion Site Based Polymorphism). The ISBP-marker is the pair of primers flanking the site of insertion of one transposable element into another, and despite the abundance of different transposable elements, ISBP could give the unique PCR product. We analyzed the composition of 5.5 % of 5BS chromosome covered by 454-reads. To perform the work related with marker anchoring we've made the selection of ISBP-markers developed on the base of 454-reads, and tested some of them for the mapping and screening of 5BS-specific BAC-library.

Methods and Algorithms: The set of 454 reads covers the 5.5 % of 5BS chromosome (39695 reads, total length 16183252 bp). The sequence composition was analyzed by RepeatMasker and search against TREP database. Also we assayed the content of specific sequences: rDNA and histone genes, centromeric and subtelomeric repeats, simple repeats and low complexity regions, mapped wheat EST (http://wheat.pw.usda.gov) and gene sequences of rice (http://rapdb.dna.affrc.go.jp). For the selection of the ISBP-markers we took 1302 markers obtained with ISBPFinder program and performed the BLAST search of amplicons against the wheat genomic sequences at NCBI database. We mapped the markers at the deletion bins of chromosome 5BS and used them for screening of BAC-library. The 5BS chromosome-specific BAC library contains 43776 clones (15-fold chromosome coverage). For fast and efficient PCR screening we performed the pooling of the library.

*Result*s: The known repeats families account for 47 % of the 454 reads, also we identified the matches with 5S rDNA, low complexity regions and the matches with mapped wheat ESTs. 12 % of ISBP-amplicons showed >90 % identity with published sequences; among them 4 % strongly matched with 5BS chromosome, 58 % had no data on chromosomal localization, 38 % matched with another chromosomes. The screening of 5BS BAC-library with some of selected ISBP-markers proved to be successful.

Conclusion: We performed the preliminary analysis of composition of the short arm of 5B chromosome, and showed the suitability of our method of markers selection for mapping and BAC-library screening, that will further contribute to the successful sequencing of the wheat 5B chromosome.

INFORMATION STORAGE IN NON-CODING DNA PATTERNS

Shadrin A.A.*[1, 2], Parkhomchuk D.V. [3]
[1] *Max Planck Institute for Molecular Genetics, Berlin, Germany;*
[2] *Design and Technological Institute of Digital Techniques SB RAS, Novosibirsk, Russia;*
[3] *Institute for Medical and Human Genetics, Charite, Berlin, Germany*
e-mail: inter.cm@hotmail.com
** Corresponding author*

Key words: *genetic information (GI), non-coding DNA, mutational spectrum, mutational load, optimization of GI storage*

Motivation and Aim: genes make up approximately 1.5% of human genome. Functional significance of remaining 98.5% non-coding DNA is still largely undetermined. At the same time, a number of recent studies show that the fraction of functional non-coding DNA in human genome may 10 times exceed the amount of protein-coding sequences. So we can expect that huge amount of genetic information is concentrated in non-coding regions of DNA. In contrast with exons, non-coding functional sites seem to be unstructured and mostly have low sequence conservation, so it is difficult currently to predict them both experimentally and computationally. To look into non-coding functionality we should first answer the question: how genetic information can be reliably stored in variable sequence patterns? In this work we put the purpose to design a model of well-defined genetic information storage in variable sequence patterns.

Methods and Algorithms: a source of inspiration for our model was the analogy between the process of sending discrete data over noisy channel and transmitting of DNA sequences through generations with mutational errors. Thus information theory (IT) was used as a mathematical basis of the model. The main challenge for such application is how to quantify the biological value of a sequence. We suggest that Schneider's definition of genetic information [1] allows quantifying this value meaningfully reflecting the degree of site conservation hence its biological value.

Results: using this approach we constructed a model of reliable genetic information storage in fuzzy sequence patterns. The cornerstones of the model are:

- GI can be stored in probabilistic patterns, with arbitrary low conservation.
- Equilibrium for a site is defined by acceptable probabilities of its alleles in a population.
- The law of GI conservation: selection maintains a pattern with constant GI through removal of some alleles.

Applying in this framework the principle of least mutational load we found that there is the optimal way of storing biological information in probabilistic patterns. The optimum predicts characteristic trends in patterns compositions which are observable in genomic data analysis. It also reveals an important evolutionary role of mutational biases in minimization of mutation load. We showed that functional patterns may bear significant amount of information having relatively low sequence conservation where about half of observed mutations are compensating "positive" mutations.

Conclusion: Further development and adaptation of proposed model can promote an improvement of computational methods of sequence pattern predictions to provide genome-wide functionality annotation. The theory allows to reinterpret some fundamental concepts, such as the neutral theory, cost-of-selection dilemma and others.

References:

1. Schneider, T.D et al. (1986) Information content of binding sites on nucleotide sequences. Journal of Molecular Biology **188**, 415-431.

NextGen SEQUENCING REVEALS EXTENSIVE RNA EDITING IN PLASMACYTOID DENDRITIC AND OTHER PRIMARY CELLS

Sharma A.[1], Alomair L.[1], Doyle K.[1], Sikaroodi M.[3], Cherepanova A.[4], Laktionov P.P.[4], Birerdinc A.[1, 2], Gillevet P. [3], Baranova A.V.*[1, 2, 3, 5]

[1] School of Systems Biology, George Mason University, Fairfax VA 22030; [2]Betty and Guy Beatty Center for Integrated Research, Inova Health System, Falls Church, VA 22042;[3] Microbiome Analysis Center, George Mason University, Manassas, VA 20100; [4]Institute of Chemical Biology and Fundamental Medicine, SB RAS, Novosibirsk, Russia; [5]Research Centre for Medical Genetics, Moscow, Russia

e-mail: abaranov@gmu.edu

** Corresponding author*

Key words: *TLR receptors, RNA editing, 454 technology*

The alteration of the nucleotide sequence of RNA, termed RNA editing (RNAE) in mammals is mainly the conversion of adenosine to inosine which is translated as if it were guanosine. This reaction is catalyzed by so-called ADARs (adenosine deaminases that act on RNA) enzymes. ADARs editing is predominant in the CNS where it is essential for correct functioning of certain neurons. However important RNAE is, only a handful of physiologically important target genes for RNAE have been confirmed in humans.

In our previous study, extensive RNAE of the A→ G (ADAR) type was revealed in human mRNA encoding Toll-like receptor TLR7 that plays a fundamental role in pathogen recognition and activation of innate immunity in plasmacytoid dendritic cells (PDCs), capable of on-demand production of type I interferons and TNF-α. RNAE of TLR7 mRNA was confirmed by confirmed by re-sequencing of TLR7 gene in the DNA prepared from the same individual' PDCs preparations. PDCs represent only 0.4 % of total peripheral blood monocytes (PBMC), and it is not surprising that theywere overlooked by other investigators of the RNAE phenomenon.

Editing of TLR7 mRNA introduce non-random changes to the coding sequence of its LRR domains responsible for molecular recognition and change its amino acid sequence. TLR7 editing was observed at high level close to 100%, edited sites result in three to four amino acid changes per individual RNA and differ betweem molecules. TLR7 mRNA editing was sample specific, i.e. it was present in PDCs of some individuals, and absent in others, thus indicating that RNA editing in PDCs is probably an inducible event and a druggable target. TLR7 editing negatively correlated with an ability of of PDCs to respond to stimulation with CpG-A. Absence of the mutations in the DNA of the same individuals was. 454 sequencing demonstrates that RNA editing is not limited to TLR7, but also occurs in TLR9 and other TLR mRNAs extracted from PDCs and the extent of RNAE differ between healthy donors. RNAE of TLRs could be responsible for mutant phenotype of their Pathogen-Associated Molecular Patterns (PAMPs)-recognizing domain, and thus individual susceptibility to common and emerging infections in otherwise healthy subjects.

Futher study will justify if the levels of TLR editing is a lifetime characteristic of the individual or it is a temporary condition reflect external influence, whether TLR are edited in DCs and influences the antigen-presenting function of DCs, how ADAR enzymes activities regulated and whether an antagonistic interaction between ADAR and RNAi machineries is pertinent to PDCs functioning.

"PROMOTER ISLANDS" AS GENOMIC REGIONS WITH QUENCHED TRANSCRIPTION

Shavkunov K.S.[1, 2], Tutukina M.N. [1, 2], Masulis I.S. [1, 2], Panyukov V.V. [3], Kiselev S.S.[1, 2], Deev A.A. [4], Ozoline O.N.*[1, 2]
[1] *Institute of Cell Biophysics, RAS;*
[2] *Pushchino State Institute of Natural Sciences;*
[3] *Institute of Mathematical Problems of Biology RAS;*
[4] *Institute of Theoretical and Experimental Biophysics RAS, Pushchino, Moscow Region, Russia*
e-mail: ozoline@icb.psn.ru, ozoline@rambler.ru
** Corresponding author*

Key words: *bacterial genomes*, *promoter islands*

Motivation and Aim: Promoter islands (PIs) were defined in the genome of *E.coli* on the basis of a high density of the transcription start points, predicted by the promoter finder PlatProm, and high ability to associate with RNA polymerase (RNAP) *in vivo* [1]. However microarray data testified their extremely low transcription activity. In this study we evaluate transcription efficiency of PIs using RNA-seq data [2] and verify their ability to form transcription-competent complexes *in vitro* and *in vivo*.

Methods and Algorithms: The genomic DNA of *E.coli* MG1655 (U00096.2, NCBI) was used for scanning and comparative analysis. The pattern of RNAP binding on the genomic DNA was assessed using ChIP-on-chip data [3]. The RNA-seq data [2] were analyzed by the RNAMatcher software. Permanganate footprinting was performed as described previously [1].

Results and Conclusion: (a) Permanganate footprinting performed for 19 out of 78 DNA fragments containing PIs, testified their ability to form 1-7 transcriptional bubbles *in vitro* and 12 PIs undergo transition into the "open" state *in vivo*. (b) At least 3 PIs initiated transcription of the reporter gene(s) if integrated into the plasmid pET28b-eGFPm-Cherry. However (c) the number of RNAs registered by RNA-seq technique [2] was less than might be expected from the RNAP binding efficiency, if any (Fig. 1). We therefore conclude that PIs are transcriptionally competent but their activity within the bacterial genome is specifically quenched as it often takes place for "alien" DNA.

Availability: RNAMatcher, which links sequence reads to the genomic coordinates by taking into account their multiple occurrence, is available by request. *Acknowledgements:* Grants of Russian Foundation for Basic Research (10-04-01218) and Russian Ministry of Education and Science are highly acknowledged.

References:

1. K.Shavkunov et al. (2009) Gains and unexpected lessons from genome-scale promoter mapping, Nucl. Acids Res. 37, 4919-4931.
2. R.Raghavan et al. (2011) Genome-wide detection of novel regulatory RNAs in *E. coli*, Genome Res., 21, 1487 - 1497.
3. N.B. Reppas et al. (2006) The transition between transcriptional initiation and elongation in E. coli is highly variable and often rate limiting, Mol. Cell, 24, 747-757.
4. T.Carver et al. (2009) DNAPlotter: circular and linear interactive genome visualization, Bioinformatics, 25, 119-120.

SEARCHING AND CLASSIFICATION OF BINDING SITES OF SIGMA FACTORS OF *CLOSTRIDIUM DIFFICILE*

Shelyakin P.V.
Institute for Information Transmission Problems RAS, Moscow, Russia
e-mail: f.serval@gmail.com, gelfand@iitp.ru

Annotation: The aim of this study is to identify genes that are regulated by different sigma factors and to obtain positional weight matrices for the promoters of Clostridia difficile 630.

Problem statement: We aim to sonstruct positional weight matrices for each sigma factor based on our knowledge of the transcriptional start points for 1500 genes of Clostridia difficile 630 [1], on the genes that potentially encode 22 sigma factors and on some genes that are known to be regulated by these factors. We then want to use those matrices to obtain information about other genes that are regulated by each sigma factor.

Results and Discussion: To construct positional weight matrices we used two methods. If sufficient experimental data on genes regulated by a certain sigma factor was available, we built the matrix using SignalX [2]. If the data were not extensive enough, we used matrices of Bacillus subtilis from the DBTBS database[3] .

Using Genome Explorer [2] we searched for genes regulated by each sigma factor in the genome of the *C. difficile* strain, for which all transcriptional start points were known, and in two genomes of other *C. difficile* strains. In the first strain we searched within the range of 40 bp from transcriptional start points; in the other two we searched in the area of 100 bp before the starting point of translation.

Having obtained three sets of genes, we then selected only genes that were preceded by candidate promoters in at least two strains. We constructed multiple alignment of their upstream nucleotide sequences in Muscle [4], after which we used SignalX to obtain positional weight matrices and Weblogo [5] to visualize the consensus sequences.

This is joint work with M. Gelfand.

References:

1. Marc Monot, Caroline Boursaux-Eude, Marie Thibonnier, David Vallenet, Ivan Moszer, Claudine Medigue, Isa Martin-Verstraete and Bruno Dupuy. *"Reannotation of the genome sequence of Clostridium difficile strain 630".* J Med Microbiol 2011 vol. 60 no. 8 1193-1199
2. Mironov A., Vinokurova N.P., Gelfand M.S., "Software for bacterial genome analysis", Mol. Biol., 2000
3. Sierro N., Makita Y., de Hoon M.J.L. and Nakai K. *"DBTBS: a database of transcriptional regulation in Bacillus subtilis containing upstream intergenic conservation information."* Nucleic Acids Res. 2008, 36 (Database issue):D93-D96; doi:10.1093/nar/gkm910
4. Edgar, R.C. (2004) *"MUSCLE: multiple sequence alignment with high accuracy and high throughput"* .Nucleic Acids Res. 32(5):1792-1797.
5. Crooks GE, Hon G, Chandonia JM, Brenner SE *WebLogo: A sequence logo generator*, Genome
6. Research, 14:1188-1190, (2004).

SECONDARY STRUCTURE OF RNA MAY CONSTRAIN INTRON EVOLUTION

Sherbakov D.Y.*, Darikova Y.A.
Limnological Institute of SB RAS, Irkutsk, Russia
e-mail: sherb@lin.irk.ru
** Corresponding author*

Key words: *intron, evolution, secondary structure, regulatory elements*

Motivation and Aim: Regulation of gene expression at the posttranscriptional level recently has attracted serious attention due to discovery of whole new class of regulatory interactions mediated by non-coding RNAs and products of their processing (1, 2). One of the least studied non-coding RNAs of potential regulatory importance are the intronic ones, their fate after the excision is generally ignored. However, it is well established that some introns undergo evolution in clearly non-neutral manner, which justifies further efforts in search for intron-contained regulatory elements (i.e. 3, 4, 5).

The present work aims at the search for introns which would evolve under rigid structural constrains such as under the condition of conservation their RNA secondary structure. The essay is based on the suggestion that in case of stem conservation the sequence evolution should proceed according to the doublet model which assume dot mutations resulting in destruction of a stem must be immediately compensated by the second substitution of the opposing base so that it becomes complementary again.

Loss of regulatory function can be the consequence if mutation is restricted only one nucleotide.

The introns evolving in accordance with doublet evolution may provide interesting candidates for further studies related to the search of particular elements whose function is depended on RNA secondary structure.

Methods and Algorithms: We used different genes introns belonging to various organisms as a model:

1) One of the introns of PFK gene coding glycolysis crucial enzyme – phosphofructokinase of gastropods mollusks from family Baicaliidae. Sequences of this intron were determined by us.

2) Nucleotide sequences of red mite (*Dermanyssus gallinae*) tropomyosin gene (TM) intron

3) Ribosomal protein (S7) first intron of *Poecilia latipinna* from fish family. Nucleotide sequences of the two last introns were taken from NCBI.

Two hypotheses were tested with Bayes ratio test as implemented in MrBayes. The first of them presumed all bases can be split into two sets one of which contained bases involved in stem double helixes, and the second one contained all of the rest ones. The first group evolves according to doublet model while the second set undergoes only single base substitutions. According to the second hypothesis the rate of substitution accumulation did not depend on secondary structure. A putative RNA conformation was modulated by RNAcofold program from Vienna RNA Secondary Structure Package (6). In program parameters 500000 replicates and 2000 burnin were set. Bayes factor was calculated by program Tracer v1.5.

Results: Introns of PFK and TM genes have shown similar results in their evolution assessment.

The Bayes ratio for them was close to 1, thus excluding statistically reliable preference to the first hypothesis. In case of protein S7 first intron this ratio was 57 that is crucial evidence supporting the first hypothesis assuming the evolution according to the doublet model. This suggests the functional capacity in its structure and consistent with other authors data about some ribosomal protein' introns involvement in coordination of these proteins synthesis.

Conclusion: Testing of two hypotheses has allowed of finding of intron evolving according to doublet model. It gives a possibility to reference this intron to potential regulatory element.

Acknowledgements: This work was partially supported by the RFBR grant 09-04-00972-a.

CYTOKINE PROFILE AND CIRCULATING DNA IN THE BLOOD OF PATIENTS WITH TICK-BORNE BORRELIOSIS

Shkoda O.S.*, Chikova E.D., Fomenko N.V., Laktionov P.P., Rykova E.Y.
Institute of Chemical Biology and Fundamental Medicine SB RAS, Novosibirsk, Russia
e-mail:olgashkoda@sibmail.com
** Corresponding author*

Key words: *Lyme borreliosis, circulating DNA, immune response, cytokines*

Motivation and Aim: Tick-borne borreliosises are widespread, characterized by high disease incidence and clinical presentations polymorphism. Their study is of high social importance due to the possibility of Lime disease chronic form development accompanied by autoimmune complications. Extracellular circulating DNA (cirDNA) were shown to be present in blood plasma in low concentrations normally, whereas their concentration has been increased in cancer and autoimmune patients blood. The aim of the study was to reveal the association of immune response development and cirDNA changes in Lime disease patients.

Methods and Algorithms: Blood samples were taken from 60 healthy subjects (HS) and 61 patients with a tick bite, which were hospitalized into Novosibirsk City Infectious Clinics №1 with signs of infection disease diagnosed as Lime disease. Blood was fractionated into plasma and cells, the cell-surface-bound cirDNA (csb-cirDNA) fraction was obtained by successive treatment of cells with PBS/EDTA and trypsin solutions. Anti-borrelia IgG and IgM antibodies titers and cytokine concentrations (IL-10, IL-6, IL-4, αTNF, γIFN) were estimated using ELISA Kits. The cirDNA concentration was evaluated using was measured by quantitative real-time PCR specific for LINE-1 repetitive elements.

Results: Reliable increase is shown of the csb-cirDNA concentration in blood from patients with Lime disease compared with healthy subjects (98 versus 13 ng/ml, Mann-Whitney U test, $p<0,0001$). Csb-cirDNA and plasma cirDNA concentration increase is associated with absence of unspecific inflammatory reaction (erythema) development (195 and 19 ng/ml in patients without erythema versus 72 and 7 ng/ml in patients with erythema, Mann-Whitney U test, $p<0,05$). Patients with Lime disease demonstrated increase of pro-inflammatory cytokine concentration (IL-6, γ-IFN, α-TNF) compared with healthy subjects (14 versus 0,1; 19 versus 1; 3 versus 0,8 pg/ml, respectively, Mann-Whitney U test, $p<0,05$). Otherwise, anti-inflammatory cytokine concentration (IL-10 and IL-4) was decreased in Lime disease patients blood compared with healthy subjects (5 versus 10; 1 versus 2 pg/ml respectively, Mann-Whitney U test, $p<0,05$).

Conclusion: The study has shown a reliable association of the cirDNA concentration increase with pro-inflammatory cytokine secretion and with the absence of unspecific inflammatory reaction during Lime disease development.

Availability: Obtained data demonstrate the cirDNA further study availability as the potential complementary factor, applied for diagnostics, and possibly, for participating in the Lime disease pathogenesis.

Acknowlegements: This work was supported by grants from RFBR № 11-04-01066-a, RAS Program “Fundamental Sciences for Medicine” № 5.25.

PATTERNS OF miRNA BINDING SITES LOCATION IN 3`UTRS OF HUMAN TRANSCRIPTS

Shtokalo D.N.*[1, 3], Saik O.V. [2], St. Laurent G.C. III[3], Kel A. [4]
[1] *A.P. Ershov Institute of Informatics Systems SB RAS, Novosibirsk, Russia;*
[2] *Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia;*
[3] *St.Laurent Institute, Providence, RI, USA;*
[4] *Genexplain GmbH, Wolfenbuttel, Germany*
e-mail: shtokalod@gmail.com
** Corresponding author*

Key words: *miRNA binding sites co-location, miRNA binding sites module*

Motivation and Aim: Recently, a new additional post-transcriptional regulation level of extensive competing relations of miRNA binding for both coding and non-coding mRNAs has been identified [1]. Study of miRNA functioning features became a very important problem considering the role of miRNAs in different biological processes including diseases. Preferred co-location of miRNA binding sites in 3`UTRs may denote the synergy of their functions [1]. This may be important for the drug development based on the miRNA interference. We aimed to study the potential miRNA binding sites co-location features in 3`-UTRs of human transcripts.

Methods and Algorithms: Data for conserved miRNA families predicted binding sites in 3'UTRs of human transcripts were obtained from TargetScan database (http://www.targetscan.org). The comparative analysis of real distribution of miRNA binding sites and simulated uniform distribution obtained by multiple shuffling (at each iteration two randomly selected genes exchange their randomly selected miRNA binding sites) was undertaken. To infer subtle structure of miRNA binding sites co-location we analyzed sites co-occurrence within the specific template window using Composite Module Analyst program [2]. This template window of 300 nucleotides contained a pair of miRNA sites within 50 nucleotides distance plus a single miRNA site. The test gene set (http://www.net2drug.info/index.html) consisted of 733 genes down-regulated in cancer cell lines after treating by a substance RITA activating p53 and 1401 control non-changed genes (gene expression was measured with microarray technology).

Results: The comparative analysis of real and simulated uniform distribution of predicted miRNA binding sites revealed the tendency for sites to be located on per miRNA specific transcripts and to avoid unspecific. So called *coefficient of specificity* was calculated for each miRNA. Analysis of subtle sites co-location in 3'UTRs of human transcripts revealed so-called *miRNA binding sites modules* M1 and M2 that are 300 nt windows where M1 comprise miR-300/381/539-3p, miR-129-5p/129ab-5p, miR-410/344de/344b-1-3p and M2 comprise miR-26ab/1297/4465, miR-495/1192, miR-374ab miRNA families binding sites. The logical construction M1 OR M2 occurs significantly more frequently in 3'UTRs of down-regulated genes comparing to non-changed genes with p-value $8.6 * 10^{-7}$ estimated by hypergeometric distribution.

Conclusion: The statistical analysis of predicted miRNA binding sites positions in 3'-UTRs of human transcripts illustrates the phenomenon of miRNA prefered binding to its specific set of transcripts and avoiding other unspecific transcripts. Also we show the presence of a subtile structure of miRNA binding sites co-location in 3'UTR of transcripts. This findings could be important for understanding the fundamental mechanisms of miRNA-target interactions, as well as for drug development based on miRNA interference.

References:

1. P. Sumazin et al. (2011) An extensive microRNA-mediated network of RNA-RNA interactions regulates established oncogenic pathways in glioblastoma, *Cell,* 147(2): 370-81.
2. T. Waleev et al. (2006) Composite Module Analyst: identification of transcription factor binding site combinations using genetic algorithm, *Nucleic Acids Res.,* 34: W541-W545.

PROTEASOMAL GENES GENOTYPE-SEX INTERACTIONS IN HUMAN POPULATIONS AND IN ASSOCIATION WITH COMPLEX DISEASES

Sjakste T.G.*[1], Paramonova N. [1], Lunin R. [1], Limeza S. [2], Sugoka O. [1], Trapina I. [1], Rumba-Rozenfelde I. [2]

[1] *Institute of Biology University of Latvia, Salaspils, LV2169, Latvia;*

[2] *Faculty of Medicine, University of Latvia, Sarlotes 1a, Riga, LV1001, Latvia*

e-mail: tanja@email.lubi.edu.lv

** Corresponding author*

Key words: *PSMA6, PSMC6, PSMA3, juvenile idiopathic arthritis, diabetes, obesity, proteasome, chromosome 14, genotype-sex interaction*

Motivation and Aim: The ubiquitin-proteasome system (UPS) potentially could affect the complex disease pathogenesis, prognosis and treatment efficiency. The UPS deregulation implicated in several human pathologies can depend on the structural variations in the genes encoding the proteasome subunits. Aim of the current study was to analyse the 14q proteasomal genes genetic diversity within and between human populations and to study the associations between genes variations and juvenile idiopathic arthritis (JIA), obesity (OB) and Type 1 *diabetes mellitus* (T1DM) in children.

Methods and Algorithms: Genotyping data characterized diversity of 38 polymorphic loci belonging to 14q proteasomal genes were produced experimentally for Latvians and extracted from public available datasets for Europeans, Asians, and Africans. Multiloci haplotypes were constructed and genetic diversity was compared between the populations. Six SNPs belonging to the *PSMB5*, *PSMA6*, *PSMC6*, and *PSMA3* genes were genotyped for association with JIA, OB and T1DM. Possible functionality of the genetic variations was analysed *in silico*.

Results: In Asians versus Europeans and Africans the allele and genotype presentation found to be opposite at the rs23480071 (*PSMA3*), and significantly different at the rs7143346 (*KIAA0391*), rs1048990 and rs17597267 (*PSMA6*), rs2295826 and rs2295827 (*PSMC6*). The rs2277459 (*PSMA6*) appear to be highly variable only in Africans. The combination from both minor alleles in the rs1048990-rs23480071 haplotype was found to be frequent in Asians (more than 20%) and rare in both Europeans and Africans (less than 5 %). The haplotype rs2295826-rs2295827-rs23480071 composed from rare alleles only appears to be unique for Asians.

The rs2277460 and rs23480071 heterozygous genotype was much more frequent in JIA, OB and T1DM patients than in controls (more than 50 % vs 35 %). For OB the association became stronger when only patients with family history were taken into account (OR = 0.32, CI = 0.17ч0.6). Double rs2277460/ rs2348071 and rs2295826/ rs2348071 heterozygotes were found to be significantly ($P < 0.001$) more frequent in JIA showing the JIA subtypes specificity and genotype-sex interaction.

Conclusion: Population and/or disease specific spectrum and level of the proteasomal genes variations may play a significant adaptive/destructive role in humans. We suggest that the *PSMA6*/*PSMA3* single/multiloci heterozygocity is a potential risk factor for complex diseases.

FAMILY OF KCTD PROTEINS: STRUCTURAL AND FUNCTIONAL PECULIARITIES

Skoblov M.Yu.*[1, 2], Marakhonov A.V. [1], Baranova A.V. [1, 3]
[1] *Federal State Budgetary Institution "Research Centre for Medical Genetics" under the Russian Academy of Medical Sciences, 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*
e-mail: mskoblov@generesearch.ru
** Corresponding author*

Key words: *KCNRG, KCTD proteins, function, tumor suppressor*

Motivation and Aim: Previously in our laboratory the novel potential tumor suppressor gene *KCNRG* had been discovered during the study of the cause of the B-cell chronic lymphocytic leukemia. The protein appeared to contain a single conserved domain, T1 potassium channel tetramerization domain. T1 domain is required for the protein interaction of proteins voltage-gated potassium channels subunit. KCNRG has been shown to negatively regulate potassium currents *in vitro* as well as suppress the proliferation and activate the apoptosis in cancer cell lines. It is known that activation and proliferation of lymphocytes depends on potassium channels. It was supposed to investigate an opportunity of an induction of apoptosis in B-CLL cells by the restoration of potassium channels inhibition with the help of low-molecular weight compounds. We also carried out i*n silico* whole-genome searching which revealed the whole family of only T1 domain-containing proteins in human genome called KCTD family. The main goal of this study was to functionally characterize the family of KCTD proteins.

Results: We selected 25 low molecular weight compounds for study its ability to inhibit potassium channels. Preliminary testing were performed on B-cell line Raji. Cells from 15 B-CLL patients isolated and maintained in culture. Ten most perspective candidates were chosen out of tested panel for further investigation. We also analyzed the participation of KCTD proteins in different cellular pathways. KCTD family showed to be a diversified group. Moreover the reported data for the interaction of KCTD family with other proteins support the hypothesis of heterogeneous functions of different KCTD proteins.

Conclusion: It is generally accepted that presence of the conserved domain in protein sequence would predominantly determine the function of the protein. This study stresses that, after initial *in silico* analysis, experimental verification of the molecular function is warranted.

References:

1. B. Biderman et al. (2010). Inhibition of potassium currents as a pharmacologic target for investigation in chronic lymphocytic leukemia. *Drug News & Perspectives*. **23** (10):625-31.
2. A Birerdinc et al. (2010). Pro-apoptotic and antiproliferative activity of human KCNRG, a putative tumor suppressor in 13q14 region. *Tumour Biol*. **31** (1):33-45.

SEARCH OF PLASMA PROTEIN BIOMARKERS FOR SCHIZOPHRENIA

Smirnova L.P.*[1], Koval V.V. [2], Loginova L.V. [1], Fedorova O.S.[2], Ivanova S.A.[1]
[1] *Mental Health Research Institute SB RAMSci, Tomsk, Russia;*
[2] *Institute of Chemical Biology and Fundamental Medicine SB RAS, Novosibirsk, Russia*
** e-mail: lpsmirnova@yandex.ru*

Key words: *schizophrenia, biomarkers, proteomic analysis*

Motivation and Aim. Schizophrenia is a complex mental disorder with fairly high level of heritability. Pathogenesis of schizophrenia is still unclear but disturbances of protein metabolism in schizophrenia are known. Nevertheless, protein marker, inherent in only this illness, still has not been detected. Objective is the proteomic analysis of blood plasma in patients with schizophrenia and healthy persons.

Methods. Object of investigation was the blood of 10 healthy persons and 16 patients with schizophrenia. Patients were under therapy at clinics of Mental Health Research Institute SB RAMSci, Tomsk. Diagnosis was conducted according to current classification ICD-10. Plasma proteins were separated with gel-electrophoresis, digested by trypsin, and analyzed by MALDI-TOF mass-spectrometry (Autoflex II, Bruker Daltonics). The proteins were identified using Mascot software (Matrix Science).

Results. Our study reveals that protein of metabotropic glutamate receptor (mGlurR6) is detected in plasma of patients with schizophrenia. Long activation of metabotropic glutamate receptors results in reinforcement of NMDA-dependent generation of active forms of oxygen what entails damage of receptors. Research of microchips has revealed decrease of expression of regulator of transmission of signal in synapses of G-protein-4 (RGS4) in schizophrenia. RGS4 is a negative regulator of receptors connected with G-proteins, including metabotropic glutamate receptors and plays an important role development of nervous system. Thus, obtained results allow to suppose that glutamatergic synapses are the basic place of action in pathogenesis of schizophrenia.

Support by Grant (N5) from SB RAS for collaborative research.

PROTEOMICS AND METABOLOMICS OF THE RAT LENS: ANALYSIS OF AGE AND CATARACT-SPECIFIC CHANGES

Snytnikova O.A.*[1,2], Kopylova L.V.[1,2], Cherepanov I.V.[1,2], Duzhak T.G.[1,2], Kolosova N.G.[3], Sagdeev R.Z.[1], Tsentalovich Y.P. [1,2]
[1] *International Tomography Center SB RAS, Novosibirsk, Russia;*
[2] *Novosibirsk State University, Novosibirsk, Russia;*
[3] *Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia*
e-mail: koa@tomo.nsc.ru
** Corresponding author*

Key words: *cataract, lens proteomics, post-translational modifications, rats*

Motivation and Aim: During the last ten years a big surgical breakthrough was made in the curing of the eye diseases, but cataract is still the most frequent cause of impairment and loss of vision in elderly people. The developed cataract is characterized by numerous post-translational modifications (PTMs) of the major lens proteins, crystallins. The transparency of the eye lens strongly depends on the crystallin solubility and structure, while the solubility is mediated by protein modifications accumulating with age. At present, it is not completely clear which modifications are cataract-specific, and which are just a part of the normal maturation and aging processes. The vast majority of experimental data on the biochemical content of cataractous human lenses corresponds to lenses with developed cataract surgically removed from patient eyes. The studies of early stages of human cataract are limited. One approach used for studies of etiology and pathogenesis of human diseases and for development of new methods for their treatment is the use of biological models. Recent studies have shown that the OXYS rat strain meets the main requirements for the model of senile cataract. The purpose of this study was to determine the age-related and the cataract-specific changes in the crystallin composition in lenses of accelerated-senescence OXYS (cataract model) and Wistar (control) rats.

Methods and Algorithms: The water-soluble (WS) and urea-soluble (US) fractions of the lens proteins were separated; the identity and relative abundance of each crystallin were determined by 2-DE and MALDI-TOF/TOF mass spectrometry.

Results: This report provides the data on the proteomic and metabolomic analysis of two rat strains of different ages: OXYS and Wistar. The interstrain differences in the crystallin compositions appear at 3 months of age. One of the most pronounced effects is the insolubilization of gamma crystallins, and this process proceeds faster in OXYS rat lenses. The main established PTMs are oxidation, N-term acetylation and asparagines deamidation.

Conclusion: The major age-related changes in proteomic composition of the rat lens are insolubilization of gamma and alpha crystallins. The major PTMs which lead to significant interstrain difference between OXYS and Wistar lenses, and, presumably, have cataract-specific character, are determined.

Acknowledgements: We appreciate: RFBR projects 11-04-00143, 11-0300296, FASI state contract 14.740.11.0758 and grant № 11.G34.31.0045, grant NSh-2429.2012.3, RAS № 21.13, CCU.

LOOKING FOR MEANINGFUL SIGNS: THE EXPERIENCE WITH COMPARATIVE ANALYSIS OF NON-ALIGNED PROTEIN SEQUENCES

Sobolev B.*[1], Oparina N.Y.[1, 2], Veselovsky A.[1], Filimonov D.A.[1], Poroikov V.V. [1]
[1] Orekhovich Institute of Biomedical Chemistry of the Russian Academy of Medical Sciences, Moscow, Russia;
[2] Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow, Russia
e-mails: borissobolev-5@yandex.ru
** Corresponding author*

Motivation and aim Large protein families frequently contain evolutionary related but functionally divergent proteins. The correct identification of amino acid positions, which can be used to discriminate functional groups within the family (Group-Specific Positions, GSP) is one of the most actual problems of Bioinformatics. Numerous approaches were developed and applied for comparative analysis of proteins. All of them require multiple alignment of amino acid sequences.

Methods and algorithms We propose a new fundamentally different approach suitable for GSP selection in non-aligned proteins. We implement the local similarities estimation, classification based on the training set and calculation of the probabilities for each region of the analyzed protein to be specificity determining.

Results We have shown the applicability of developed algorithm for detection of various types inter-subfamilies functional variations, including even a weak signals associated with protein-nucleic acid and protein-protein interaction.

Conclusion We have shown the applicability of alignment-free approach for protein families analysis and group-specific positions detection.

Availability The developed software is freely available as a web service at http://195.178.207.160/spros/.

THE UNDERLYING MECHANISMS OF REPROGRAMMING OF HUMAN UMBILICAL VEIN ENDOTHELIAL CELLS (HUVEC)

Sokolov A.S.*[1], Mazur A.M. [1], Vassina E.M. [2], Prokhortchouk E.B. [1], Zhenilo S.V. [1]
[1] *Center Bioengineering RAS, Moscow, Russia;*
[2] *Vavilov Institute of general genetics RAS, Moscow, Russia*
e-mail: sokolovbiotech@gmail.com
** Corresponding author*

Key words: *IPSC, HUVEC, reprogramming, ChIP-seq*

Motivation and Aim: Induced pluripotent stem cells (iPSCs) offer immense potential for regenerative medicine and studies of disease and development. Somatic cell reprogramming involves epigenomic reconfiguration, conferring iPSCs with characteristics similar to embryonic stem (ES) cells. However, it remains unknown how complete the reestablishment of ES-cell-like chromatine marks patterns and genes expression patterns is throughout the genome. Here, we try to describe underlying mechanisms of reprogramming of human umbilical vein endothelial cells (HUVEC) by retroviral overexpression of Oct4, Sox2, cMyc.

Methods and Algorithms: To assess the degree to which a somatic cell chromatine marks and gene expression pattern is reprogrammed into an ES-cell-like state, we generated ChIP-seq analysis using antibodies to active chromatine mark - H3K4me2 and silent chromatine mark - H3K27me3 and gene expression analysis using Illumina microarrays (Illumina HumanRef-8). The principal component analysis to highlight the grouping of iPSCs and ESCs far from the starting cell type has been conducted in R . Graphs were made using the first two components. We used QuEST_2.4 software for finding regions of enrichment of chromatine marks.

Results: We performed principal component analysis using data from gene expression experiment (our expreiment and experiments available on-line - GSE20532, GSE26575, GSE28688) and ChIP-seq experiment (our expreiment and experiments available on-line - GSE26386, GSE23455). Reassuringly though, performing a principal component analysis for each reprogramming experiment available, we always saw that iPSCs are much closer to ESCs than to the starting somatic cells, based on their genome-wide transcriptional profile and genome-wide ChIP-seq profile (bivalent marks). Also we find that pluripotency genes are inactive in differentiated cells, are located in heterochromatin regions (H3K27me3), and not expressed. During reprogramming, silent pluripotency genes become activated (not all, but for Oct4 we saw enrichment for H3K4me3). At the same time, tissue-specific genes become silent and undergo the opposite processes (not for all tissue-specific genes). This suggests than in our reprogramming experiment, although close to ESCs, iPSCs contain a gene-signature that could differentiate them from ESCs in accordance to or that iPSCs and ESCs are not strictly equivalent on a transcriptome and chromatine marks levels.

Conclusion: In conclusion we proposed that our IPSCs lines have somatic memory. This feature may restrict potential of IPSCs lines for regenerative medecine.

STUDY OF PROMOTERS OF *YODA* AND *BHSA* GENES ENCODING STRESS RESPONSE PROTEINS IN *E. COLI*

Sokolov V.S.*, Likhoshvai V.A., Khlebodarova T.M., Oshchepkov D.Y., Efimov V.M., Babkin I.V.
Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
e-mail: sokovlad1@bionet.nsc.ru
** Corresponding author*

Key words: *regulation of transcription, genosensor, stress response, yodA gene E. coli, bhsA gene E. coli, mathematical model*

Motivation and Aim: The purpose of this work is study of the regulation transcription mechanisms of stress-sensitive *yodA* and *bhsA* genes in *E.coli*. The choice of these genes was based on data available in literature.
Methods and Algorithms: To solve this problem, two special plasmids were constructed. These constructs contained promoter of *yodA* or *bhsA* gene and *gfp* gene, which played a role of reporter. Bioinformatics methods such SITECON program and Hill's functions were used also.
Results: We have used *E.coli* cells with above-mentioned plasmids to observe how different toxic agents influence transcription of *gfp* under various conditions. We have predicted the transcription factors binding sites on these promoters by using SITECON program. Based on the results we have constructed the mathematical model of regulatory mechanism of *bhsA* gene. Our mathematical model of transcription regulatory mechanism of genes is tentative.
Conclusion: The constructed genosensors responses under different concentrations of toxic agents and temperatures were obtained. The potential transcription factors binding sites on promoters of genes under consideration was computed by using SITECON program. The tentative mathematical model described functioning of *bhsA* gene promoter was constructed.

COMPUTATIONAL TOOLS FOR ANALYSIS OF NEXT GENERATION SEQUENCING DATA

Solovyev V.[1], Seledtsov I.[2], Vorobyev D.[2], Kosarev P.[2], Molodsov V.[2], Okhalin N.[2]
[1] *Department of Computer Science, Royal Holloway, University of London, UK;*
[2] *Softberry Inc., 116 Radio Circle, Suite 400, Mount Kisco, NY 10549, USA*
e-mail: victor@cs.rhul.ac.uk

Next-generation sequencing (NGS) transforms today's biology research by providing fast sequencing of new genomes, genome-wide association studies (GWAS), sequencing personal genomes, variant discovery by resequencing targeted regions or whole genomes, de novo assemblies of bacterial and eukaryotic genomes, annotating the transcriptomes of cells, tissues and organisms (RNAseq), and gene discovery by metagenomics studies. To analyze next-generation sequencing data we advanced further our **OligiZip** assembler and **Transomics pipelines** that provide solutions to the following tasks: 1) de novo reconstruction of genomic sequence; 2) reconstruction of sequence with a reference genome; 3) mapping RNA-Seq data to a reference genome and identification of alternative transcripts with quantification of their abundance; 4) fast alignment of assembled contigs to the genomic sequence.

To test reconstruction of bacterial sequences we assembled genomic sequence of Methanopyrus kandleri TAG11 and Methanopyrus kandleri AV19. Solexa reads, about 6 million each for AV19 and TAG 11, were produced by sequencing lab of Harvard PCPGM. AV19 genome itself has been assembled perfectly in one contig. Time of AV19 assembling is ~ 1 min on one Linux node. **OligiZip** also has been successfully applied to assembly a model eukaryotic genome in Assemblathon 1 collaborative efforts where many research groups presented genome assembles to estimate the accuracy of various assembling software.

For RNAseq data the **Transomics** pipeline initially maps reads to the genomic sequence and identifies spliced and non-spliced reads coordinates. This information used by our FGENESH gene prediction program that includes an iterative procedure for predicting alternative splicing gene variants. We have developed a module to compute a relative abundance of predicted alternative transcripts solving a system of linear equations. The initial variant of Transomics pipeline has been successfully applied to Human, C.elegans and Drosophila data of the RGASP project. We also developed a powerful **Sequence assembling Viewer** to work with the reads data and assembling results interactively. As an example of Transomics application for identification of disease specific genes we analyzed RNAseq data for non-tumorigenic epithelial cell line and epithelial cells from infiltrating ductal carcinoma of the breast. Comparative analysis shows a set of genes that have different alternative splicing forms in these cell lines.

OligoZip, Genome Aligner and **Transomics** pipeline components and other software programs are available to run independently at www.softwberry.com or as a part of integrated environment of the **Molquest** software package that can be downloaded at www.molquest.com for Windows, MAC and Linux OS.

INTEGRATIVE CELL MODELING USING DATA INTEGRATION AND TEXT MINING APPROACHES

Sommer B.
Bio-/Medical Informatics Department, Bielefeld, Germany
e-mail: bjoern@CELLmicrocosmos.org

Key words: *Cell Modeling, Data Integration, Text Mining, Visualization*

Motivation and Aim: For understanding intracellular relationships, the prediction and/or verification of proteomic localizations are crucial tasks. Usually, a set of disease-related proteins – often correlated with a biological network – is distributed throughout different cell components. It is a time-consuming task to find and compare different potential localizations reported in a large number of different publications and other resources in the Internet.

Methods and Algorithms: The CELLmicrocosmos PathwayIntegration [1] is used to correlate different protein-associated networks with a virtual cell environment. For this purpose, two methodologies are combined:

1. Data Integration: The data warehouse DAWIS-M.D. [2] contains different life-science relevant databases which are used for the localization and the generation of KEGG-pathways [3].
2. Text mining: The ANDCell database [4] is used to use localization information obtained from PubMed abstracts.

Results: Different two-dimensional as well as three-dimensional visualization methods are used to enable a fast analysis and assignment of protein localizations based on published material.

Conclusion: The combination of different techniques and collaboration among different disciplines enables a profound localization prediction of pathway-associated data as well as disease-related protein sets.

Availability: http://Cm4.CELLmicrocosmos.org (Restricted Access until Full Release)

Acknowledgements: This work was supported in part by: BMBF Internationale Zusammenarbeit in Bildung und Forschung mit Russland, Projekt RUS 08/005; DFG Graduate College for Bioinformatics (GK635).

References:

1. B. Sommer et al. (2010) Visualization and Analysis of a Cardio Vascular Disease-and MUPP1-related Biological Network Combining Text Mining and Data Warehouse Approaches. *Journal of Integrative Bioinformatics,* **7**:148.
2. K. Hippe et al. (2010) DAWIS-MD-A Data Warehouse System for Metabolic Data. *GI Jahrestagung* **2**:720–725.
3. M. Kanehisa et al. (2012) KEGG for Integration and Interpretation of Large-scale Molecular Data Sets. *Nucleic Acids Research,* **40**:D109–D114.
4. P.S. Demenkov et al. (2008) Associative Network Discovery (AND)-the Computer System for Automated Reconstruction Networks of Associative Knowledge About Molecular-genetic Interactions. *Computational Technologies,* **13**:15–19.

THE VESICLE BUILDER – A PLUGIN FOR THE CELLmicrocosmos 2 MembraneEditor

Sommer B.*, Yan Zhou
Bio-/Medical Informatics Department, Bielefeld, Germany
e-mail: bjoern@CELLmicrocosmos.org
** Corresponding author*

Key words: *Membrane Modeling, CELLmicrocosmos2 MembraneEditor, 3D Membrane Packing Problems, Vesicle Generation*

Motivation and Aim: The CELLmicrocosmos 2.2 MembraneEditor (CmME) is a Java WebStart program to solve heterogeneous, two-and-a-half-dimensional Membrane Packing Problems [1]. It imports and exports lipids and proteins based on the PDB format [2]. Originally, CmME was developed for generating rectangular membranes which are also often used in conjunction with molecular simulations. A demanding task is the generation of lipid-based spheres or vesicles.

Methods and Algorithms: Here, the first version of a new CmME-plugin is presented, enabling the solution of three-dimensional Lipid Packing Problems. An algorithm as implemented which enables the generation of ellipsoid vesicular mono-layer or bilayer membranes.

Results: The generation of ellipsoid vesicles is now a simple task by using the new *Vesicle Builder*.

Conclusion: The new methodology extends the capabilities of the MembraneEditor to solve now additionally three-dimensional Lipid Packing Problems. The next steps will be to extend the framework of CmME to support its well-known features also for vesicles, for example: the definition of microdomains, the automatic computation of lipid densities and the semi-automatic placement of proteins.

Availability: http://Cm2.CELLmicrocosmos.org

Acknowledgements: This work was supported in part by: DFG Graduate College for Bioinformatics (GK635).

References:

1. B. Sommer et al. (2011) CELLmicrocosmos 2.2 MembraneEditor: A Modular Interactive Shape-Based Software Approach To Solve Heterogeneous Membrane Packing Problems. *Journal of Chemical Information and Modeling,* **5**:1165–1182.
2. H.M. Berman et al. (2000) The Protein Data Bank. *Nucleic Acids Research,* **28**:235–242.

THE RECURRENT HORIZONTAL TRANSFERS OF DIFFERENT TRANSPOSABLE ELEMENTS BETWEEN LEPIDOPTERA SPECIES

Sormacheva I.*, Blinov A.
Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
e-mail: sormacheva@bionet.nsc.ru
** Corresponding author*

Key words: *Tc1/mariner DNA transposons*, *CR1 non- LTR retrotransposons*, *Lepidoptera*, *phylogeny*, *horizontal transfer*

Motivation and Aim: Horizontal transfer (HT) is considered an important part of the evolutionary process of eukaryotic genomes. Most examples of HT of eukaryotic genes involve transposable elements. Although HT is well known for DNA transposons and LTR retrotransposons, non-LTR retrotransposons rarely undergo HT, and their phylogenies are largely congruent to those of their hosts. Previously we described HT of CR1-like non-LTR retrotransposons between butterflies (*Maculinea*) and moths (*Bombyx*) which occurred less than 5 million years ago and we also hypothesized that DNA transposons could be involved in HT of non-LTR retrotransposons [1]. In present study we explored the diversity of CR1 non-LTR retrotransposons and Tc1/mariner DNA transposons, as potential vectors for HT of non-LTR retrotransposons.

Methods and Algorithms: We used a combination of bioinformatical tools and molecular biology techniques to analyze Tc1/mariner DNA transposons and CR1B non-LTR retrotransposons in the Lepidoptera genomes.

Results: Our results demonstrate diverse distribution of different subgroups of transposable elements among lepidopterans. We have showed multiple cases of possible HTs of Tc1/mariner and CR1B elements between *Maculinea* and *Bombyx* genera. They occurred recently (approximately 5 MYA) and affected a group of species that diverged more than 140 MYA [2]. We also identified chimeric Tc1/mariner /CR1B sequences in *Maculinea* and *Bombyx* genomes which could represent vestiges of transposon-based linked HT of these elements between moths and butterflies.

Conclusion: We performed comprehensive analysis of two groups of transposable elements (Tc1/mariner and CR1-like elements) in Lepidoptera genomes by bioinformatical and molecular biological approaches. While we did not find strong evidence for our hypothesis of the involvement of DNA transposons in HT of non-LTR retrotransposons, we demonstrated that recurrent and/or simultaneous flow of TEs took place between distantly related moths and butterflies. The search for possible vectors, such as viruses (or virions) and intracellular parasites, seems warranted.

References:

1. O.Novikova, E.Sliwińska, V.Fet, J.Settele, A.Blinov, M.Woyciechowski. (2007) CR1 clade of non-LTR retrotransposons from Maculinea butterflies (Lepidoptera: Lycaenidae): evidence for recent horizontal transmission, BMC Evol Biol., **7**: 93.
2. M.W.Gaunt, M.A.Miles. (2002) An insect molecular clock dates the origin of the insects and accords with palaeontological and biogeographic landmarks, Mol Biol Evol., **19**:748-761.

MODULATION OF TISSUE REGENERATION WITH CHEMICALS THAT AFFECT ION CHANNELS: MODEL ORGANISM *MACROSTOMUM LIGNANO*

Sormacheva I.*[1], Berezikov E. [2]
[1] *Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia;*
[2] *Hubrecht Institute, Utrecht, the Netherlands*
e-mail: sormacheva@bionet.nsc.ru
** Corresponding author*

Key words: *Ion channels, membrane voltage, regeneration, stem cells, flatworms, Macrostomum lignano*

Motivation and Aim: Ion channels represent a diverse family of membrane-spanning proteins that allow the flow of inorganic ions across cellular membranes and thus establish membrane voltage potential. Due to their diverse biological roles, the importance of ion channels as a target class for modern drug discovery has been recognized. The bioelectrical membrane polarization, mediated by ion channels, is important in regulation of tissue regeneration processes. Several ion channel-targeting drugs, such as an antiparasitic agent ivermectin, can lead to changes in body axis formation if flatworm *Dugesia japonica* and induce regeneration of head structures in place of tail [1, 2]. In the present study we propose to use *Macrostomum lignano* as a model to evaluate effects of numerous known ion channel - targeting drugs on regeneration.

Methods and Algorithms: We have analyzed effects of ion channel - targeting drugs on regeneration of *Macrostomum lignano* with use of bioinformatical and molecular biological methods.

Results: We have demonstrated that ion channel - targeting drugs ivermectin and praziquantel change patterns of membrane polarization and concentration of intracellular calcium ions in intact flatworms *Macrostomum lignano* and during the initiation of regeneration process. Ivermectin and praziquantel treatment can induce initiation of head structures regeneration at the tail fragments in *Macrostomum lignano*, which under normal conditions cannot regenerate head regions at all [3]. Thus it appears that chemical targeting of ion channels can be a potent approach for modulation of regeneration programs and potentially can be of value for regenerative medicine.

Conclusion: Macrostomum lignano can be widely used as a model for drug discovery for treatment of different "channelopathies", including diabetes, cystic fibrosis, hypertension, cardiac arrhytmia and a variety of neuronal disorders.

References:

1. T.Nogi, D. Zhang, J.D.Chan, J.S.Marchant. (2009) A novel biological activity of praziquantel requiring voltageoperated Ca^{2+} channel beta subunits: subversion of flatworm regenerative polarity, PLoS Negl. Trop. Dis., **3**:e464.
2. W.S.Beane, J.Morokuma, D.S.Adams, M.Levin. (2011) A chemical genetics approach reveals H,K-ATPase-mediated membrane voltage is required for planarian head regeneration, Chem Biol., **18**:77-89.
3. B.Egger, P.Ladurner, K.Nimeth, R.Gschwentner, R.Rieger. (2006) The regeneration capacity of the flatworm Macrostomum lignano - on repeated regeneration, rejuvenation, and the minimal size needed for regeneration, Dev Genes Evol., **216**:565- 575.

ELECTROSTATIC PROPERTIES OF T7 DNA PROMOTERS

Sorokin A.A.*[1], Beskaravainy P.M. [1, 2], Osypov A.A.[1], Kamzolova S.G. [1]
[1] *Institute of Cell Biophysics RAS, Pushchino, Russia;*
[2] *Institute Theoretical and Experimental Biophysics RAS, Pushchino, Russia*
e-mail: lptolik@gmail.com
** Corresponding author*

Key words: *T7 phage, promoter, electrostatics of DNA*

Motivation and Aim: Physical properties of genome DNA are well recognized factor in the processes of transcription regulation. Bacteriophage is unique organism in that its expression controls by two different RNA polymerases, comparison of promoters controlled by both enzymes gives us an opportunity of better understanding of transcription initiation mechanisms.

Methods and Algorithms: Here, electrostatic potential distribution of the complete sequence of T7 DNA genome was calculated by Coulombic method [1,2]. Eight σ^{70}-specific promoters interacting with *E.coli* RNA polymerase ($E\sigma^{70}$) and 17 promoters recognized by T7 phage specific RNA polymerase were localized in T7 DNA electrostatic map. Comparative analysis of electrostatic properties for promoter and nonpromoter sites in T7 DNA was carried out.

Results: Nonpromoter sites are characterized by more smoothed electrostatic profiles as compared with complicated rich in details patterns of promoter sequences. It should be noted that electrostatic profiles of promoters recognized by two RNA polymerases differ in their size and design. Characteristic electrostatic patterns for σ^{70}- and T7-specific promoters embrace ~200 b.p. (-150 -- +50) and ~50 b.p. (-35 -- +15) correspondingly. The diference in the size of the patterns closely falls within the range of DNA contacting sites for the two enzymes which is to be accounted for by the difference in their size.

Conclusion: Thus, electrostatic potential distribution around DNA provides effective means for identification of promoter sites in T7 genome and their differentiation by different RNA polymerases.

References:

1. S G Kamzolova, V S Sivozhelezov, A A Sorokin, T R Dzhelyadin, N N Ivanova, R V Polozov. (2000) RNA polymerase-promoter recognition. Specific features of electrostatic potential of “early” T4 phage DNA promoters. *J. Biomol. Struct. Dyn.* **18** p. 325-334
2. R V Polozov, T R Dzhelyadin, A A Sorokin, N N Ivanova, V S Sivozhelezov, S G Kamzolova. (1999) Electrostatic potentials of DNA. Comparative analysis of promoter and nonpromoter nucleotide sequences. *J. Biomol. Struct. Dyn.* **16** (6) p. 1135 -- 43.

PHYSICAL PROPERTIES OF T7 NATIVE PROMOTERS DNA: CONTRIBUTION OF DNA ELECTROSTATICS AND DUPLEX STABILITY TO PROMOTER EFFICIENCY

Sorokin A.A.*, Dzhelyadin T.R., Kamzolova S.G.
Institute of Cell Biophysics RAS, Pushchino, Russia
e-mail: lptolik@gmail.com
** Corresponding author*

Key words: *T7 phage, promoter, DNA thermostability, electrostatics of DNA*

Motivation and Aim: Native T7 promoters despite high homology in their nucleotide sequence show significant variation in transcription initiation rate. There are evidences that the difference could be explained by the analysis of physical properties of promoter DNA.

Methods and Algorithms: Distribution of electrostatic potential around the DNA of complete T7 genome was calculated by Coulombic method [2]. Stress-induced Duplex Destabilized Sites of 1000 bp fragmets arount all 17 native promoters recognized by T7 RNA-polymerase was calculated via WebSIDD server [3,4].

Results: Native T7 promoters could be devided into two classes. All class III promoters have exact 23-bp sequence around transcription start site. Class II, which deviates from 23-bp consensus sequence and is weaker than class III promoters. We have shown that difference in promoter strength could be explained by presence of highly unstable DNA upstream of class III promoters and more stable duplex around class II promoters. At the same time duplex stability could not be treated a promoter determinant because a lot of unstable sites are not recognized by RNAP, while class II promoters that have stable DNA structure are able to initiate transcription however with lower efficiency.

Conclusion: Role of DNA stability profile in efficiency of transcription initiation by T7 RNAP has been shown and use DNA electrostatics, duplex stability and primary structure for prediction of promoter efficiency is discussed.

References:

1. S G Kamzolova, V S Sivozhelezov, A A Sorokin, T R Dzhelyadin, N N Ivanova, R V Polozov. (2000) RNA polymerase-promoter recognition. Specific features of electrostatic potential of “early” T4 phage DNA promoters. *J. Biomol. Struct. Dyn.* **18** p. 325-334
2. R V Polozov, T R Dzhelyadin, A A Sorokin, N N Ivanova, V S Sivozhelezov, S G Kamzolova. (1999) Electrostatic potentials of DNA. Comparative analysis of promoter and nonpromoter nucleotide sequences. *J. Biomol. Struct. Dyn.* **16** (6) p. 1135 -- 43
3. C.-P. Bi, C.J. Benham. (2004) WebSIDD: Server for Prediction of the Stress-induced Duplex Destabilized Sites in Superhelical DNA, *Bioinformatics*, **20**, 1477-1479.
4. WebSIDD calculation site

INVESTIGATION OF VOLE *Nanog* REGULATORY REGION

Sorokin M.A.*, Elisaphenko E.A.
Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
e-mail: sorokin@bionet.nsc.ru
** Corresponding author*

The *Nanog* gene is one of the key regulators of mammalian pluripotency. In this work, we cloned and sequenced the regulatory region of East European vole *Nanog* sized 4.5 kb upstream of transcription start site. Multiple alignment of the vole *Nanog* regulatory region and the homologous sequences of other mammals allowed determination of the most conservative regions which can participate in regulation of the gene in vole. To find out a role of DNA methylation in the vole *Nanog* transcriptional regulation, a methylation status of CpG-rich fragments of the *Nanog* regulatory region in vole embryonic and extraembryonic lineages was studied. Some potential transcription factor binding sites were predicted in the vole *Nanog* regulatory region by the MathInspector software. Analysis of published data on participation of transcription factors in regulation of the mouse *Nanog* and ChIP-seq and ChIP-on-chip experiments on protein binding in the mouse *Nanog* locus was carried on. Based on these mouse data positions of potential regulatory elements in vole *Nanog* upstream region was narrowed. Then the regulatory activity of these particular regions was analyzed by the reporter construct assay. Thus, a general model of the vole *Nanog* transcriptional regulation was suggested.

VARIABLE PATTERNING IN DROSOPHILA EMBRYOS DUE TO BASINS OF ATTRACTION IN UNDERLYING GENE REGULATORY DYNAMICS

Spirov A.V.*[1, 2], Holloway D.M. [3]
[1] *The Sechenov Institute of Evolutionary Physiology and Biochemistry, Saint-Petersburg, Russia;*
[2] *Computer Science, State University of New York at Stony Brook, Stony Brook, NY, USA;*
[3] *Mathematics Department, British Columbia Institute of Technology, Burnaby, B.C., Canada*
e-mail: alexander.spirov@gmail.com
** Corresponding author*

Key words: *pattern formation, gene network modeling, dynamical systems, bifurcation, phase diagrams*

Motivation and Aim: Fruit flies (*Drosophila*) are model organisms for studying spatial pattern formation in animals. In the first few hours of development, a network of interacting genes forms expression patterns which determine the body plan. Data shows that wild-type (WT) development is remarkably robust, with various initial trajectories canalizing to an attracting state [Surkova et al., 2008]. Dynamical systems analysis of a core nonlinear model of the segmentation gene network has shown how this WT stability can arise as a trajectory through phase space [Manu et al., 2009ab]. Using 'coarse-grained' reaction-diffusion modeling, in which gene-gene interactions are simplified to single signed connections, we can study the robustness in the zygotic segmentation network via dynamical systems analysis and computations.

Methods and Algorithms: Protein expression for the 4 early segmentation genes (*gap* class) is modeled using the gene circuit framework [Manu et al., 2009ab], producing AP concentration patterns. The dynamics of an N variable gene circuit can be represented by behavior in an N-dimensional concentration or phase space. Time-varying solutions follow trajectories in the phase space; stable solutions are given by fixed points. In [Manu et al., 2009ab], the phase space was mapped numerically and fixed points classified according to their stability. The positions of the fixed points and their stability properties determine the stability of a general time varying solution of the gene circuit, including bifurcation points between neighboring basins of attraction. In the present work, we computed a number of trajectories with different initial conditions to test stability and the reduction of variability.

Results and Conclusion: We have shown here how mutation of gene-gene interactions can lead the Drosophila segmentation gene network to a bifurcation point, at which natural maternal variability can push embryos into neighboring basins of attraction. Such variable expressivity or incomplete penetrance is observed in nature, but the causes have been elusive. Our work suggests a dynamical basis, in which a weak mutation takes the system to a bifurcation point, and the variable outcomes are a manifestation of natural variability in upstream control; i.e. the mutation removes the robustness of the gene network to maternal variability. Understanding the model components and parameters which produce the experimental pattern perturbations allows us to map the 'near-WT' phase space, and by so doing, to create a detailed understanding of the biological regulatory dynamics used in body formation.

Acknowledgements: This work supported by NIH grant R01-GM072022.

DISCOVERY AND VALIDATION OF SERUM BIOMARKERS FOR MONITORING OF DISEASE PROGRESSION AND THERAPEUTIC RESPONSE IN DUCHENNE MUSCULAR DYSTROPHY

Spitali P.[1], Hiller M. [1], Nadarajah V. [1], Martin C. [1], Oonk S. [1], van der Burgt Y. [2], den Dunnen J.T. [1], van Ommen G.J. [1], Aartsma-Rus A. [1], 't Hoen P.A.C. [1]
[1] Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands;
[2] Department of Parasitology, Leiden University Medical Center, Leiden, The Netherlands

Motivation and Aim: Muscular dystrophies represent a subset of neuromuscular disorders among which Duchenne Muscular Dystrophy (DMD) is the most common and the most severe. The two main focuses in our lab are the development of therapeutic approaches to treat muscular dystrophies and the discovery/validation of biomarkers able to non-invasively track disease progression in DMD patients.

Results and Discussion: DMD is caused by out of frame mutations in the *DMD* gene. We and others showed that antisense oligonucleotides (AONs) mediated exon skipping is able to restore dystrophin expression in DMD muscle cells. We used gene expression profiling to find transcriptional biomarkers able to describe the muscle phenotype in several dystrophic mouse models and we could prove that exon skipping mediated dystrophin rescue is able to restore the gene expression signature in dystrophin KO mice. The developed AONs have been licensed to Prosensa/GSK and are currently tested in a phase 3 clinical trial in Duchenne patients. Given the quick clinical development of the AONs for DMD, we further studied the possibility to find non-invasive biomarkers in the serum of DMD patients. We developed a targeted and a non-targeted approach which enabled us to find putative biomarkers which we are currently in the validation phase.

DYNAMICAL AND STRUCTURAL ANALYSIS OF AN APOPTOSIS NETWORK IN HEPATITIS C

Stepanenko I.L.*, Smirnova O.G.
Institute of Cytology & Genetics, SB RAS, Novosibirsk, Russia
e-mail: stepan@bionet.nsc.ru;
** Corresponding author*

Key words: *apoptosis, boolean model, gene network, topology, hepatitis C virus, signal transduction pathway, proliferation*

Motivation and Aim: Chronic hepatitis C virus (HCV) infection is a major risk factor for liver disease progression, and may lead to cirrhosis and hepatocellular carcinoma (HCC). Apoptosis plays a significant role in the pathogenesis of hepatitis C and is a host defense mechanism against viral infection and hepatocarcinogenesis. Studies of local network topology and dynamics can be used to investigate as well to predict novel candidate therapeutic targets for HCC. In the present study we reconstruct the apoptosis gene network and analyze their topology and dynamical properties.

Methods and Algorithms: The GeneNet technology was used for description the regulation of apoptosis by HCV proteins. Boolean model that assume only two states (ON or OFF) for each component, was used to describe the behavior of the gene network. The Boolean rule for each node is determined based on the nature of interactions between the nodes. This rule can be expressed using the logical operators AND, OR and NOT.

Results: HCV-regulated gene network in apoptosis was reconstructed on the basis of the data extracted from 273 experimental papers. It includes 36 genes, 121 proteins, and 280 molecular interactions. The GeneNet Hepatitis C (apoptosis) scheme is available {{http://wwwmgs.bionet.nsc.ru/mgs/gnw/genenet/}}. We found that apoptosis gene network is organized in a modular, hierarchical manner. A few nodes with a very large number of links, which are called hubs, are represented by the central transcription factors, NF-κB and p53. HCV proteins modulate various signal transduction pathways binding directly with the key signal molecules and transcription factors. At least four of the 10 HCV proteins, namely core, NS3, NS5A and NS5B play roles in several potentially oncogenic pathways. HCV viral proteins could modulate therapeutic responses by altering host cell microRNA (miRs) expression.

Using a synchronous boolean model we have simulated the network dynamics. Dynamical analysis allowed us to enumerate the stationary states and to find the key network regulators.

By microarray analysis, several potential gene markers of HCV-associated liver disease and more than 100 genes with the altered expression levels were identified. The information about hepatitis C virus-induced genes, transcription regulation, regulatory regions, and transcription factor binding sites is collected in TRRD database, section HCV-TRRD.

Conclusion: This model can be useful to test the coherence of experimental data and to hypothesize gene interactions that remain to be discovered. The application of mathematical methods that quantitatively describe the properties of apoptosis gene network is crucial for studying complex diseases such as hepatitis C and cancer.

Availability: http://wwwmgs.bionet.nsc.ru/mgs/gnw/genenet/viewer/index.shtml
http://wwwmgs.bionet.nsc.ru/mgs/papers/stepanenko/hcv-trrd/

Acknowledgements: The work was supported by grant EU-FP7 SYSPATHO № 260429.

OSBORN LAW OF ADAPTIVE RADIATION AS A BACKGROUND OF AROMORPHOSES

Suslov V.V.
Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
e-mail: valya@bionet.nsc.ru

Key words: *Osborne law of adaptive radiation, Haldane's dilemma, aromorphosis, stress*

Motivation. Aromorphosis scenarios are morphologically substantiated. However, they do not explain the regularities noted by Osborn. An invasion to a pessimal habitat is performed (i) by small populations; (ii) parallel in different subtaxa and/or in parts of the geographic range of the taxon[1]; (iii) regardless of the value of preadaptations (a taxon poor in preadapations disperses, and rich does not)[2]; (iv) preadaptations arise parallel (may be in close time). Preadaptation is necessary but not sufficient. An acceleration of adaptive evolution in small populations is restricted by Haldane's dilemma. Neutralism favors neither invasion nor parallelisms [1].

Results. Osborn explained the regularities of invasions by selection for hyperactivity, but he failed to explain the causes of this selection and advantages favoring hyperactive individuals. Rapid evolution of a small population, parallelisms, and success of invasions into different habitats, loosely linked to preadaptation euadaptivity, can be associated with *par force* evolution (see the corresponding thesis). Selection involves few genes[3], prolonged cross-resistance phase of stress allows brief excursions to various pessimal habitats[4], and parallelisms in evolution of stress-associated gene networks is supported at least at three levels: common genes[5], common functions, and a limited kit of operative 3D structures of proteins[6] [1].

Acknowledgements: RFBR 10-04-01310, Integration Projects of Presidium RAS 6.8, 30.29, No.28 «BOE», Science Schools 5278.2012.4.

References:

1. V.V. Suslov (2012) Par force evolution as a mechanism of rapid adaptation, *Paleontology Journal*, (Mosk), (In press).

[1] It may be illustrated by infrequent but systematic findings of dispersing individuals. Since 1870s, rare flights of the cattle egret (*Bubulcus ibis*) to South America have been recorded. A stable population formed in 1930s. A prominent example is the mastering of perching on trees by pigeons that occurred parallel (even synchronously) and independently in various parts of their synanthropic range within few generations.

[2] An illustrative example is the colonization of waterless areas by modern fish species. Although eupreadaptations (lung, swimbladder, intestinal, labyrinth, etc respiration) are common, their possessors are mainly aquatic species. They either involuntarily quit water or survive dry periods in torpor. Some species are normally active outside water: those with inadaptive simple air respiration systems (life forms like eels or Periophthalmus) or even species without such systems able to arrest respiration like grunion, flying fish, and hatchetfish [1].

[3] The rest of the variability *at the time of a stress factor action* is neutralized, or suppressed by the group stress effect according to Gorban, or has no vital significance: Phenotype destabilization, stress-induced mutagenesis, and population genostasis are tested by selection *long after the action of the stress factor*.

[4] Thus, the structure of an econiche according to Hutchinson should be supplemented by a stress margin exceeding the limits of the basic econiche but accessible for a subpopulation of stress-tolerant individuals, alternatively, for the whole population if the proportion of such individuals is above a certain threshold.

[5] In particular, a change in the expression of the common gene fraction in prokaryotes has been found to respond to various stress factor groups: (i) hyper- and hypoosmosis, low and high temperatures and pressures; (ii) radioresistance and drought tolerance; (iii) hypoxia, starvation, and xenobiotics [1].

[6] Optionally, there may be a common stressor (Darwinian convergence).

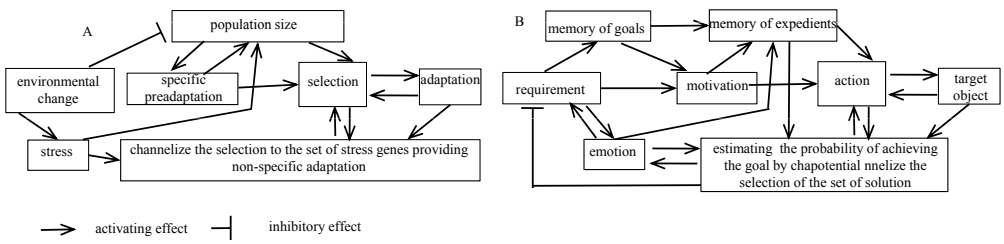
STRESS AND HOMEOMORPHY OF ADAPTATION MECHANISMS

Suslov V.V.
Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
e-mail: valya@bionet.nsc.ru

Key words: *norm of reaction, non-specific adaptations, stress, emotion*

Motivation. Analysis of interactions of specific and non-specific adaptations is important open problem of evolution researches. During invasion of species to pessimal or unstable habitats (environment changes faster than alternation of generations), search of a pool of mutations for pre-adaptations is complicated. Alternative scenario – temporary acceleration of mutability or destabilization of phenotype – increase selection cost threatening population by degeneration. Similar difficulties meet going back to Shmalhausen and Turesson scenarios of a genocopying of modifications within wide norm of reaction (NR) if the share of preadaptive modifications in NR of taxon is small[1] or the preadaptive modifications ontogenesis is blocked environmentally[2] [2].

Results. The scenarios described lack feedback link between phenotype estimating pessimality of environment and evolving genome. Such link could be explained by stress if start of stress response of organism and selection in the population are consequences of the same environment change (Fig. A). We revealed examples of this scenario realization and limits of its application. The scheme interactions of specific and non-specific adaptations suggested is close to scheme of behavior act by Simonov [3]. Combining of non-specific emotion (emotion doesn't bear information on requirement) and specific motivation gives estimation of the probability of the aim achievement (Fig. B).



Acknowledgements: RFBR 10-04-01310, IP Presidium RAS 6.8, 30.29, No.28 «BOE», НШ-5278.2012.4.

References:

1. G.F. Gause (1940) On the importance of adaptability for natural selection, *Journ. Gen.Biol.* 1(1): 105-120. (in Russ.)
2. V.V. Suslov (2012) Par force evolution as a mechanism of rapid adaptation, *Paleontology Journal*, (Mosk), (In press).
3. P.V. Simonov. Emotional brain. Mosk., 1981. 215 p. (in Russ.)

[1] Thus, at infusoria wide NR gave adaptation only to a "familiar" factor – often met in phylogenesis. Selection by an "unfamiliar" factor was endured by pre-adaptive mutants [1].

[2] Environmentally block is best studied on behavior modifications. Experience of the felines breeding in captivity revealed wide NR on social behavior. But in the nature only lions at top of a food pyramid of savannas are social. Here, acting as an environment factor, they block a sociality of felines, close on an adaptive zone. Elimination of lions by the humans caused fast socialization of cheetahs [2].

TENDER FOR A ECOLOGICAL NICHE AS CONDITION OF THE ARO(ALLO)MORPHOSES

Suslov V.V.
Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
e-mail: valya@bionet.nsc.ru

Key words: *evolution, aromorphosis, allomorphosis, stress, Osborne invasions*

Motivation. Par force evolution (see the corresponding thesis) able provide successive invasion to new ecological niche (license), but not sufficiently long taxon habitat in such a niche. It is in agree with paleontology data: after aromorphosis long period of "fitting" of morphophysiological system is observed, i.e. temps of evolution are growing, but its variance decreasing[1] [1]. Hence taxon-invader in new niche should be vulnerable in competition with multiple population of the aboriginal taxon that is not in a stasis condition. Competitive exclusion by Gause was described for case of effective pre-adaptation formed long before, but not during time of invasion to new habitat.

Results. We formalize conception of "tender to niche" – situation when an estimate of taxon adaptivity in the given niche is possible disregarding its population size[2]. Tender to niche for related species is hybridization+introgression of genes (*i*); tender for niche for unrelated species is break of succession by guild of cenophobes that allows to intercept resources from aboriginal species without the direct competition (*ii*); tender for niche via biotope – accumulation of non-utilized mortmass and metabolites make biotope more favourable for invader (*iii*); *par force* tender – canalization of evolution, limiting space of mutation selection by more perspective to adaptivity mutations, for example selection by optimization of stress-response or strengthen other ways of non-specific adaptation (i*v*). Adaptations, leading to the tender to niche, are altruistic, that provide temporary conservation of the tender (in limit even after death of all cenopopulation of claimers (most striking example – after invaders' death mortmass preserve their tender in bioptope). Subpopulation of stress-resistant species (*iv*) provide repeatability, but not always success (here is altruism) of invasions completing tenders *i*-*iii*, that do not give repeatability[3].

Acknowledgements: RFBR 10-04-01310, IP Presidium RAS 6.8, 30.29, No.28 «BOE», НШ-5278.2012.4.

References:

1. A.S. Rautian, Paleontology as a source of information about the patterns and factors of evolution, *Modern paleontology*. v. 2. Mosk., 1988. 76-118. (in Russ.).

[1] It means temps of taxon formation in such a new niche accelerate, but rank of new taxons is lower

[2] Traditional estimate of invader adaptivity as fitness [3] is possible only in average, but often is not applicable during invasion to new habitat. For example, if invader feed in new niche, but reproduce in ancestral niche, fitness conception may lead to paradoxes: hippopotamus (may give birth, but not feed in water) is adapted to water license, but platypus and seals are not (not give birth and offspring not live in water). But in mammalian fauna of Australia namely egg-laying had open this license for platypus species.

[3] Except of independent parallel Osborne invasions (see the corresponding thesis) repeatability provide stress-resistant settling and persisting stages of life cycle. Stress-adaptation often is resistance to own non-utilized metabolites. Selection by stress-resistance may shift time of reproduction, making easy cross-species hybridization.

SINGLE AND PAIR CHANGE POINTS IN GENE SEQUENCES

Suvorova Y.M.*, Korotkov E.V.
Center of Bioengineering RAS, Moscow, Russia
e-mail: suvorovay@gmail.com
** Corresponding author*

Key words: *change points, coding sequence, triplet periodicity*

Motivation and Aim: It is widely known that nucleotide composition is not absolutely homogeneous within genetic sequences and that this heterogeneity could not be explained just by random fluctuation. From this heterogeneity arise a task of segmentation of the sequence to a set of homogeneous parts. Many different methods have been developed for the purpose of detecting change points in DNA sequences. Most approaches were proposed for genome sequences. But there is some heterogeneity on a gene level too. We focused on the single and pair CP in DNA coding sequence.

Methods and Algorithms: Working with coding sequences we can use their well known property: "triplet periodicity" (TP). TP is a common property of all known living organisms and also it is associated with a gene reading frame [1]. So one can find in the sequence segments within which TP is the same or nearly the same and between which TP types are different. And the positions between these segments we call change points (CP) of TP. We developed a mathematical method to identify TP CP along a gene sequence (S). A frequency matrix is a representation of TP in a sequence. This matrix reflects the number of a base of i type ('A','C','T','G') that is in a position j of codon (j=1, 2, 3). Our measure is based on comparison these matrixes constructed for subsequences of S. We mowed a sliding pointer x along S and compared left [x-L, x] and right [x+1, x+L] subsequences. For each x we have varied the length of considering subsequences (L=60...600 symbols), searching such length that would maximize the difference value F. We select the point at which difference reaches its maximum and then checked its statistical significance. In case of paired CP search we used additionally measure of matrix similarity based on the Bessel function to confirm homogeneity of TP between points.

Results and Conclusion: Using the method we have analyzed genes from the KEGG-48 database (~$4 \cdot 10^6$ sequences). The total number of significant results was 311 221(with false positives ~5%) [2]. We found that only about 7% of the result set could be explained by repetitive or low-complexity regions. We suppose that the TP changes may indicate a fusion of genes or domains in the region. We performed BLAST analysis to find potential ancestral genes for the parts of genes with TP changes. As a result we found that in 131323 cases have proper similarities (in different genes) for one or both parts (i.e. left and/or right from x). We also identified 2700 genes with pair CP among 66936 genes from 17 bacterial genomes.

Availability: The executable file of CP search program is available by request from authors.

References:

1. F.E.Frenkel, E.V.Korotkov. (2008) Classification analysis of triplet periodicity in protein-coding regions of genes, *Gene*, **421**:.52-60.
2. Y.M.Suvorova, V.M. Rudenko, E.V.Korotkov. (2012) Detection change points of triplet periodicity of gene. *Gene.* **491**: 58-64.

DESCRIPTION OF A LATERAL ROOT DEVELOPMENT IN TERMS OF THE GROWTH TENSOR

Szymanowska-Pułka J.
Department of Biophysics and Plant Morphogenesis US, Katowice, Poland
e-mail: jsp@us.edu.pl

Key words: *lateral root formation, growth rates, plant organ modelling, Arabidopsis*

Motivation and Aim: Growth of lateral root, like of other plant organs, is anisotropic, so three principal directions of growth can be distinguished within it. This allows for applying tensorial tools to description of the organ growth [1]. The main objective of this study was to describe lateral root formation in *Arabidopsis thaliana* and to analyze a potential role of the principal directions in this process.

Methods and Algorithms: Features of the cell pattern of developing lateral roots in *Arabidopsis* were studied and a tensorial model for growth and division of cells [2], [3] specified for this case. An unsteady character of the growth field [4] of the organ was assumed.

Results: Some significant features observed in lateral roots, such as oblique cell walls present at subsequent developmental stages, prove that the principal directions of growth play an important role in the cell pattern formation. In simulations based on a model for growth two variants for cell division have been applied: in variant I cells divide in one of the principal direction, in variant II a division wall is perpendicular to the nearest existing cell wall. From simulation sequences of the organ development have been obtained, in which the features present in anatomical sections can be recognized mostly in variant I. Based on literature data the formation of particular tissue types have been reconstructed in the virtual lateral root.

Conclusion: 1) The model for growth and division of cells [2], based on the growth tensor concept [1] have been applied to the lateral root formation in *Arabidopsis*. 2) Characteristic features of the cell pattern observed in empirical data can be identified in apices obtained from simulation. 3) A role of principal directions in the pattern formation has been confirmed in the simulation: the variant of cell division in accordance to these directions have given more realistic result.

Availability: N/A

Acknowledgements: This work was supported by the Polish Ministry of Science and Higher Education [grant number N N303333936].

References:

1. Z.Hejnowicz, J.A.Romberger. (1984) Growth tensor of plant organs, *Journal of Theoretical Biology,* **110:** 93-114.
2. J.Nakielski. (2008) The tensor-based model for growth and cell divisions of the root apex. I. The significance of principal directions. *Planta* **228:** 179-189.
3. J.Szymanowska-Pułka. (2007) Application of a changing field of growth rates to a description of root apex formation. *Journal of Theoretical Biology* **247:** 650-656.
4. J.Szymanowska-Pułka, J.Nakielski. (2010) The tensor-based model for growth and cell divisions of the root apex. II. Lateral root formation. *Planta* **232**: 1207-1218.

NOVEL APPROACH FOR IDENTIFICATION OF DNA-BINDING PROTEINS OF BLOOD CELL SURFACE

Tamkovich S.N.*[1,2], Duzhak T.G. [3], Starikov A.V.[4], Vlassov V.V.[1], Laktionov P.P. [1]
[1] *Institute of Chemical Biology and Fundamental Medicine SB RAS, Novosibirsk, Russia;*
[2] *Novosibirsk State University, Novosibirsk, Russia;*
[3] *International Tomografic Center SB RAS, Novosibirsk, Russia;*
[4] *Novosibirsk Regional Oncology Dispensary, Novosibirsk, Russia*
e-mail: s.tamk@niboch.nsc.ru

** Corresponding author*

Key words: *cell-surface-bound DNA, DNA-binding proteins, blood cells, proteomics, breast cancer*

Motivation and Aim: It was shown, that endogenous extracellular DNA (cirDNA) circulate in blood and other biological fluids as free DNA, in the complexes with biopolymers (nucleosomes, etc.) and being bound with surface of blood cells. Cell-surface bound DNA may be composed of free cirDNA molecules or supramolecular complexes of cirDNA. As far as circulation biological effects and elimination of cirDNA from blood are mediated by their interaction with cells, binding of cirDNA with cellular surface is of enormous importance. Earlier we found that in blood of healthy donors more then 90% cirDNA are bound to the surface of blood cells (csbDNA), whereas in blood of breast cancer patients less then 10% cirDNA are bound to cell surface [1]. Redistribution of cirDNA in blood of breast cancer patients allows to execute a comparative study of circulating nucleoprotein complexes bound with surface of blood cells in healthy women and BC patients.

Methods and Algorithms: The proteins mediating cirDNA binding with cell surface were identified by mass-spectroscopy after isolation of nucleoprotein complexes eluted from blood cell surface by DNA-affinity chromatography. In the pilot experiments the following proteins were identified: C3 complement component, haptoglobin, fibrinogen, serum albumin, apolipoprotein and 59,5 kDa histidine-riched glycoprotein.

Results and conclusion: Identification of proteins from circulated nucleoprotein complexes will allow to reveal the characteristic features of generation and circulation cirDNA in blood of breast cancer patients. The data obtain are demanded for isolation of tumor-specific DNA, identification of the mechanisms of binding and penetration of cirDNA into cells, provide a new data regarding molecular pathology of breast cancer and could potentially help to identified new protein markers of breast cancer.

References:

1. S.N. Tamkovich et al. (2005) Level of extracellular nucleic acids bound with blood cell surface in diagnostics of breast cancer, *Mol. Med. (Russian),* **2:** 46-50.

HIGH THROUGHPUT SSR CHARACTERIZATION AND LOCUS DEVELOPMENT FROM NEXT GEN SEQUENCING DATA

Tchourbanov A.
Beijing Institute of Genomics (BIG), Building G, No.7 Beitucheng West Road, Chaoyang District, Beijing 100029, P.R.China
e-mail: alexander@big.ac.cn

Key words: *Microsatellites, hidden Markov model, genotyping*

Motivation and Aim: Microsatellites are among the most useful genetic markers in population biology. High-throughput sequencing of microsatellite-enriched libraries dramatically expedites the traditional process of screening recombinant libraries for microsatellite markers. However, sorting through millions of reads to distill high quality polymorphic markers requires special algorithms tailored to mollify sequencing errors in locus reconstruction, distinguish paralogous loci, rarify raw reads originating from the same amplicon, and sort out various artificial fragments resulting from recombination or concatenation of auxiliary adapters. Existing programs warrant improvement.

Results: We describe a microsatellite prediction framework named HighSSR for microsatellite genotyping based on high throughput sequencing. We demonstrate the utility of HighSSR in comparison to Roche gsAssembler on two Roche 454 GSFLX runs. The majority of the HighSSR-assembled loci were reliably mapped against model organism reference genomes. HighSSR demultiplexes pooled libraries, assesses locus polymorphism, and implements Primer3 for the design of PCR primers flanking polymorphic microsatellite loci. Hypothetically, the framework could also be used for SSR genotyping entirely based on high throughput sequencing without using traditional PCR amplification and sizing.

Availability: http://code.google.com/p/highssr/

Acknowledgements This work was supported by National Science Foundation DBI-0821806 and DEB-0817033.

PREVENTING COMMON HEREDITARY DISORDERS THROUGH TIME-SEPARATED TWINNING

Tchourbanov A.
Beijing Institute of Genomics (BIG), Building G, No.7 Beitucheng West Road, Chaoyang District, Beijing 100029, P.R.China
e-mail: alexander@big.ac.cn

Key words: *Artificial twinning, in vitro fertilization, complex diseases, Mendelian disorders, prevention, preimplantation genetic screening*

Motivation and Aim: Biomedical advances have led to a relaxation of natural selection in the human population in developed countries. In the absence of strong purifying selection, spontaneous and frequently deleterious mutations tend to accumulate in the human genome and gradually increase the genetic load; that is, the number of potentially lethal genes in the gene pool. It is not possible to assess directly the negative impact of the genetic load on modern society because it is influenced by many factors such as constantly changing environmental conditions and continuously improving medical care. However, if modern medicine loses its effectiveness, significantly higher than normal mortality is expected before equilibrium with the environment is re-established. Recent advances in *in vitro* fertilization (IVF) combined with artificial twinning and improved embryo cryoconservation offer the possibility of preventing significant accumulation of genetic load and reducing the incidence of hereditary disorders.

Discussion: Many complex diseases such as type 1 and 2 diabetes, autism, bipolar disorder, allergies, Alzheimer disease, and some cancers show significantly higher concordance in monozygotic (MZ) twins than in fraternal twins (dizygotic, DZ) or parent-child pairs, suggesting their etiology is strongly influenced by genetics. Preventing these diseases based on genetic data alone is frequently impossible due to the complex interplay between genetic and environmental factors. We hypothesize that the incidence of complex diseases could be significantly reduced in the future through a strategy based on time-separated twinning. This strategy involves the collection and fertilization of human oocytes followed by several rounds of artificial twinning. If preimplantation genetic screening (PGS) reports no aneuploidy or known Mendelian disorders, one of the MZ siblings would be implanted and the remaining embryos cryoconserved. Once the health of the adult MZ sibling(s) is established, subsequent surrogate parenthood with the cryoconserved twins could substantially lower the incidence of hereditary disorders with Mendelian or complex etiology.

Summary: The proposed method of artificial twinning has the potential to alleviate suffering and reduce the negative social impact induced by dysgenic effects associated with known and unknown genetic factors. Time-separated twinning could deliver more accurate health predictions for children of surrogate parents compared to estimates based on the health records of sperm and egg donors.

Acknowledgements: Supported by grant 2011Y1SA09 from the Chinese Academy of Sciences Fellowship for Young International Scientists and by grant 31150110466 from the National Natural Science Foundation of China (NSFC).

CLUSTERING OF *E. COLI* PROMOTER ELECTROSTATIC PROFILES

Temlyakova E.A.*, Kamzolova S.G., Sorokin A.A.
Institute of Cell Biophysics RAS, Pushchino, Russia
e-mail: evgenia.teml@gmail.com
** Corresponding author*

Key words: *promoter, electrostatics of DNA, clustering*

Motivation and Aim: The profile of the electrostatic potential around *E. coli* promoter DNA was shown to differ from the profile around the coding regions [1]. Wide distribution of values of electrostatic potential in the promoter and considerable overlap with distribution of electrostatic potential in coding area makes it difficult to use this characteristic for prediction of transcriptional start site. Proper classification of electrostatic patterns in the promoter regions could shed an additional light to the mechanism of promoter-polymerase interaction. It could also bring a powerful tool for prediction of promoter position, its biological functions and strength.

Methods and Algorithms: We have applied Ward's agglomerative hierarchical clustering [2] to the collection of more than 350 experimentally identified *E. coli* promoters. They were grouped into clusters by similarity of their electrostatic profile in the area of core promoter (-45 -- 0 bp) or upstream region (-90 -- -55 bp) and the stability of the obtained groups were tested [3]. Genes whose expression is controlled by promoters from the collection were grouped into sets according to identified stable promoter clusters. Then each gene set was tested for overrepresentation of biological function with the gene set enrichment analysis [4].

Results: Two and three stable clusters were obtained within the core promoter and the upstream region electrostatic profiles correspondently. It was shown that promoters controlling similar biological processes were clustered together, for example most of genes involved in aromatic compound metabolism are expressed from promoters belonging to one cluster.

Conclusion: Identification of functionally related *E. coli* promoters in the corresponding clusters obtained on the basis of their electrostatic potential profiles may indicate the presence of common regulatory mechanisms based upon physical properties of their promoter DNA. Extraction of specific electrostatic elements participating in those mechanisms would not only improve our understanding of the process of prokaryotic transcription initiation but would also help in prediction and functional annotation of regulatory elements in bacterial genomes.

Acknowledgements: This work was supported by RFBR grant 11-04-01436-a.

References:

1. R V Polozov, T R Dzhelyadin, A A Sorokin, N N Ivanova, V S Sivozhelezov, S G Kamzolova. (1999) Electrostatic potentials of DNA. Comparative analysis of promoter and nonpromoter nucleotide sequences. *J. Biomol. Struct. Dyn.* **16**(6): 1135 - 1143.
2. L. Kaufman,and P.J. Rousseeuw. (1990). Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York.
3. T. I. Simpson, J. D. Armstrong, A. P. Jarman. (2010). Merged consensus clustering to assess and improve class discovery with microarray data. *BMC Bioinformatics,* **11**: 590
4. 4. S.Falcon, R.Gentleman. (2007) Using GOstats to test gene lists for GO term association, *Bioinformatics*, **23**(2): 257-258.

DEVELOPMENT AND APPLICATION OF THE GENOMIC CONTROL METHODS FOR GENOME-WIDE ASSOCIATION ANALYSIS USING NON-ADDITIVE MODELS

Tsepilov Y.A.*[1], Read J.[2], Strauch K.[2], Axenovich T.I.[1], Aulchenko Y.S.[1]
[1]*Institute of Cytology and Genetics SD RAS, Novosibirsk, Russia*
[2]*Helmholtz Centre Munich, Germany*
e-mail: drosophila.simulans@gmaol.com
** Corresponding author*

Key words: *genome-wide association analysis, genomic control, non-additive models of inheritance*

Motivation and Aim: Genome-wide association (GWA) analysis is a powerful tool for mapping genes of complex traits. GWA analysis assumes that the samples are randomly ascertained from the same genetic population. In this case, the phenotypes of individuals correlate only with the genotypes of loci that are involved in the control of the trait. In reality, correlations are also caused by confounders that are in correlation both with the phenotype and the genotypes of various loci. Among most prominent factors, genetic (sub)structure of the sample can work as such confounder, leading to false-positive genetic associations. However, in the framework of GWA analysis the availability of genotypes of large set of markers allows to adjust the results of the analysis using the genomic control (GC) method. At present stage, the GC is formulated and implemented for additive models. The aim of this work was to develop, validate and implement methods of genomic control for various models of inheritance.

Results: We have derived analytical expressions for adjustment factors for association test statistics that depend on the allele marker frequency and the parameters describing the population-genetic structure of sample data. We also proposed a new method of correcting the test statistics by the polynomial depending on the allele frequency. Obtained expressions, procedures of parameter estimation and procedures for the correction of the analysis results are implemented as a computer program written using R language. The software developed was used to characterize the statistical properties of the implemented methods - described earlier, and introduced us. Methods were applied to study real data.

Conclusion: Results for non-additive genome-wide association analysis cannot be adjusted using GC methods proposed for the additive model of inheritance. We proposed several methods and strategies that can be used for correction of non-additive genome-wide association analysis results. The use of non-additive models of inheritance can improve the power of GWA analysis and help discovering new loci involved in control of complex traits.

MOLECULAR EVOLUTION OF PROTEINS BELONGING TO AUXIN BIOSYNTHESIS GENE NETWORK IN PLANTS

Turnaev I.A.*, Akberdin I.R., Mironova V.V., Omelyanchuk N.A., Afonnikov D.A.
Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
e-mail: turn@bionet.nsc.ru
** Corresponding author*

Key words: *auxin biosynthesis, plants, proteins, paralogs*

Motivation and Aim: The aim of our study was to determine the evolution of auxin biosynthesis pathways in plants by analysis of homologous enzymes of these pathways in the lower and higher members of this kingdom.

Methods and Algorithms: To form the samples we searched for homologs of 18 plants, 2 yeast and 3 bacterial auxin biosynthesis enzymes among proteins from 15 species with fully sequenced genomes. Search for homologues was performed using BLAST, with E-value $\leq 10^{-7}$. As a result, 11 samples of proteins were found. Narrowing the number of samples up to 11 is due to the fact that there was a large number of paralogs in the initial sample. Belonging to groups of orthologs was determined by using the data from the database KEGG ORTHOLOGY.

Results and Conclusion: (1) In our experiments we found no homologues of both bacterial {Ps: trp-2 aminotrasferase} and {Bs: IPDC} and yeast {Sc: trp aminotrasferase} proteins, which suggests that these auxin biosynthetic pathways are absent or relatively rare in plants.

(2) We revealed that number of paralogs for some enzymes of auxin biosynthesis may vary greatly depending on species. There were more paralogs IGPS in *Sorghum bicolor, Oryza sativa, Brachypodium distachyon* and *Zea mays*; TSA - in *Medicago truncatula*; DOPA in rice; ASA1, ASB1 and PAI2 - in *Arabidopsis thaliana*). In opposite, there were species, which lost some paralogs, ASB1 is absent in *Medicago truncatula*, IGPS – in *Lotus japonicus* and TSB - in Zea mays). The increase in the number of paralogs may indicate the specific importance of the corresponding phases in auxin biosynthesis for these species. (3) We found that the first part of the auxin biosynthesis pathway, which occurs from chorismate to tryptophan (Normanly, 2010) is inherent to all studied groups of higher and lower plants, while the second part of the auxin biosynthesis pathway from tryptophan to auxin is absent in lower plants such as red and green algae. This part of the pathway appears first in evolution in the moss. This indicates that the auxin in ancient plant groups was synthesized only by tryptophan independent way, and tryptophan-dependent auxin biosynthesis pathway appeared in the evolution of higher plants to provide the complexity of their morphogenesis.

Acknowledgements: This work was supported in part by grants 11-04-01748-a, 11-04-01254-a, НШ-5278.2012.4 and integration projects RAN 6.8, 28 and 30.29.

References:

1. J. Normanly. (2010) Approaching Cellular and Molecular Resolution of Auxin Biosynthesis and Metabolism, *Cold Spring Harb. Perspect. Biol.,* 2**(1)**:a001594.

SEARCHING FOR REGULATORY CIRCUITS IN GENE NETWORKS

Turnaev I.A.*[1], Kalgin K.V. [2], Afonnikov D.A. [1]
[1] *Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia;*
[2] *Institute of Comput. Mathematics and Mathem. Geophysics SB RAS, Novosibirsk, Russia*
e-mail: turn@bionet.nsc.ru
** Corresponding author*

Key words: *Negative feedback loops, positive feedback loops, regulatory circuits*

Motivation and Aim: There are two important motifs in genetic networks (GN), among others, negative feedback loops (NFLs) and positive feedback loops (PFLs). NFLs maintain homeostasis in living systems, PFLs provide departure of a system from the steady state. However, the recognition of the FLs in the gene networks graph as circuits (directed closed paths in graph) is imprecise because some FLs are not circuits in the gene networks graph. For example, protein regulation of its own degradation and inhibition of enzyme activity by substrate are FLs, but they are not circuits in the gene network graph. In this work we suggest approach to overcome this problem by using modification of gene network graphs representation by adding extra edges from vertices of reactions to vertices of their substrates.

Methods and Algorithms: The input graph of gene network should be represented in Pajek format. The NFLs/ PFLs search is based on the Jonson's algorithm of finding elementary circuits in directed graph [1].

Results: The extended representation of gene network graphs allows recognizing all FLs as the circuits. Our method allows identifying three types of regulatory circuits in these networks: 1) PFLs, 2) NFLs and 3) neutral circuits, consisting of reactions without regulatory events. We implemented our approach to *Minimum cell cycle* [2] (37 vertices: 5 genes, 10 protein and protein complexes, 12 reactions, 20 regulatory events; 47 edges) and *Yeast cell cycle* [3] (55 vertices: 2 genes, 17 protein and protein complexes, 19 reactions, 19 regulatory events; 71 edges) gene networks. For example, the number of identified PFLs after graph extension increased from 10 to 15 in the *Minimum cell cycle*, and from 12 to 16 in the *Yeast cell cycle*. Similar results were obtained for NFLs. At the same time the number identified neutral circuits does not change.

Conclusion: The analysis of extended graphs allows detecting 33-39% feedback loops (PFLs + NFLs) more than in the usual gene networks representation.

Acknowledgements: This work was supported in part by grants НШ-5278.2012.4 and integration projects RAN 6.8, SB RAS integration projects 130 and 39.

References:

1. Jonson D.B., (1975) Finding all the elementary circuits of a directed graph, *SIAM J. on Comput.,* **4**, 77-84.
2. Turnaev I.I. et.al., (2007) Theoretical analysis of the evolution of molecular cell cycle control, "CPMS' 2007", Conference proceedings Nov. 16-19, 2007, Moscow, Russia, 335-338.
3. Novak B. et.al., (2001) Mathematical model of the cell division cycle of fission yeast. *Chaos*, **11**(1):277-286.

FUNCTIONAL INTERPLAY OF OVERLAPPING PROMOTERS PREDICTED WITHIN *phoR/brnQ* "PROMOTER ISLAND"

Tutukina M.N.*[1, 2], Lukyanov V.I.[3], Kiselev S.S.[1, 2], Ozoline O.N.[1, 2]
[1] *Institute of Cell Biophysics, RAS;*
[2] *Pushchino State Institute of Natural Sciences;*
[3] *Institute of Theoretical and Experimental Biophysics RAS, Pushchino, Russia e-mail: maria@icb.psn.ru*
** Corresponding author*

Key words: *bacterial genomes*, *Escherichia coli*, *aRNAs*, *promoter islands*

Motivation and Aim: Promoter islands (PIs) are specific regions in the genome of *E.coli* with high density of the transcription start points (TSP) predicted *in silico*, high ability to bind with RNA polymerase (RNAP) *in vivo,* but low transcription efficiency [1]. The goal of this study was to analyze the principles of promoter selection by RNAP within a particular PI located in *phoR/brnQ* intergenic region.

Methods and Algorithms: E.coli K12 MG1655 (U00096.2, NCBI) strain was used for all experiments. *Salmonella enterica* Ty2 genome (NC_004631) was used for comparison. Genomes were scanned using different version of PlatProm software [1]. RNAP binding was tested by band-shift assays, permanganate footprinting and ChIPon-chip data [2]. Transcription activity was checked *in vitro* (single-round), *in situ* (GFP reporter system) and *in vivo* (qRT-PCR).

Results and Conclusion: PlatProm predicted a region with high density of TSPs in the very end of *phoR* gene and intergenic region *phoR/brnQ*, while 2 kbp upstream were free from transcription signals. The same pattern was found in the *S.enterica* genome suggesting its conservation among closely related bacterial species. Band-shift assays and ChIP-on chip data indicated multipoint RNAP binding in this region, while permanganate footprinting revealed 5 open complexes. Products of *in vitro* transcription corresponded to all of them, assuming high transcriptional output. However both microarrays [2] and qRT-PCR testify only three promoters that are active *in vivo*. One of them initiates synthesis of antisense RNA to *phoR*, while 2 other can drive *brnQ* transcription. The most active one (P2) is located 170bp upstream of this gene and possesses stringent control discriminator. It gave strong product in primer extension and increased GFP fluorescence 35 fold. Previously mapped *brnQ* promoter [P1, 3] located 69bp downstream showed extremely low activity in our assays and even inhibited transcription from P2. Our data exemplify functional interplay between multiple overlapping promoters that is basically important for transcriptional control within promoter islands.

Acknowledgements: We are grateful to Russian Foundation for Basic Research (grants 10-04-01218, 12-04-01830) and Russian Ministry of Education and Science for supporting this project.

References:

1. K.Shavkunov et al. (2009) Gains and unexpected lessons from genome-scale promoter mapping, Nucl. Acids Res. 37, 4919-4931.
2. N.B. Reppas et al. (2006) The transition between transcriptional initiation and elongation in *E. coli* is highly variable and often rate limiting, Mol. Cell**,** 24, 747-757.
3. A. Mendoza-Vargas et al (2009) Genome-wide identification of transcription start sites, promoters and transcription factor binding sites in *E.coli*, PLoS One, 4 (10):e752

ON 3D RECONSTRUCTION AND LINEAGE OF *ARABIDOPSIS* EMBRYOS FROM A COLLECTION OF OBSERVATIONS OF FIXED SAMPLES BASED ON CONFOCAL MICROSCOPY

Urbain A.[1], Palauqui J.-C.[1], Nikolaev S.V.[2], Kolchanov N.A.[2], Trubuil A.*[1]
[1] *INRA, Paris, France*
[2] *Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia*
e-mail: alain.trubuil@jouy.inra.fr
** Corresponding author*

Key words: *plant embryo, development, confocal microscopy, 3D reconstruction, image processing*

Motivation and Aim Development of plant embryo beginning from the very early stages is not fully studied yet. For example, detailed description of dynamics of cellular composition of embryo, and the cells growth are unknown. Herein we propose a quantitative description of the early *Arabidopsis thaliana* development. The description is quantitative with respect to cell shapes, division planes, dynamics though time lapse data are not available.

Methods and Algorithms We have designed three algorithms. The first one is devoted to 3D reconstruction from stacks of confocal images. Cell walls are highlighted with Propidium iodide and detected with a crest line operator. This crest information is combined with watershed segmentation. The second algorithm builds a tree representation of the embryo. Each node of the tree corresponds to a set of related cells and each branch points towards a daughter cell. This representation is based on topological and geometrical rules applied on the 3D reconstructions. The third algorithm is devoted to embryos comparisons. Considering two embryos at different stages of development and their associated tree representations we propose to plug the younger embryo inside the older one. This is realized by tree explorations from the root. We use Avizo package for surface meshing and 3D visualization and matlab package for image processing.

Results 3D geometrical models of early stages of Arabidopsis embryo development were constructed in meshes representation. Different features such as volumes, neighboring structure have been computed. For several embryos, the history of divisions has been reconstructed. Growth rates could be inferred from the plug-in of embryos.

Conclusion From a collection of *Arabidopsis* embryos observed in confocal microscopy quantitative data such as volumes can be analyzed. Despite the fact that no time lapse is available some dynamics information could be inferred. At least several rounds of divisions can be guess. Of course variability should be addressed and the plug-in algorithm we propose can be helpful. All the 4D information we got can now be used for designing real models for biomechanics and transport phenomena evaluation during early plant embryo development.

Acknowledgements The work was partly supported by the RFBR-INRA grant 11-04-91397 «Acquisition of bilateral symmetry in embryos of dicotyledonous plants».

DATA MINING TOOL FOR ANALYSIS OF REGULATORY REGIONS OF GENES: INTEGRATION OF ExpertDiscovery AND UGENE

Vaskin Y.Y. *[1], Vityaev E.E. [2], Khomicheva I.V. [3]
[1] *Novosibirsk State University, "UniPro" Company, Novosibirsk, Russia;*
[2] *Istitute of Mathematics SB RAS, Novosisbrsk, Russia;*
[3] *Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia*
e-mail: vaskin90@gmail.com
** Corresponding author*

Key words: *Complex signals, relation data mining, bioinformatics, UGENE, recognition, gene regulatory regions*

Motivation and Aim: Rapid development in the field of sequencing technology caused accumulation of huge amount of data. To analyses this data effectively, we need to use modern bioinformatics tools. The ExpertDiscovery system is designed to analyze one of the most important parts of eukaryotic genome – gene regulatory regions. The task of extraction of the hierarchical structure of the regulatory regions is an actual and poorly studied problem.

The first goal of the project is development of the scope of the ExpertDiscovery system for more convenient, fast and efficient recognition of the complex signals. Due to this development, experts will have a tool, which is in the context of interacting modules of UGENE. The second goal of the project is a using and presentation of the system on real data.

Methods and Algorithms: The ExpertDiscovery system which is integrated into UGENE package performs data mining of sequences using semantic probabilistic inference based on a training set. Rules, which are discovered by the system, are complex signals that reflect the hierarchical structure of the region. Then we can recognize signals on any sequence, taking into account different statistics which are displayed as a few numerical parameters and graphs. The key feature of the system is a possibility of combining of results, which are obtained with other methods, to extract more complex rules.

Results: A convenient tool for analysis of regulatory regions has been integrated into UGENE package, providing a good possibility of automation of expert's work. All the operations are performed within the environment of one tool which speeds up productivity of work.

Conclusion: An original approach to data integration of different methods and knowledge of experts was developed. During the study, significant rules, which cannot be extracted automatically with other similar methods, have been found.

Availability: www.ugene.unipro.ru. Free open-source software. GPLv2 license.

VARIABILITY OF GENE EXPRESSION IN MOUSE BRAIN DEPENDS ON PREDICTED TBP-AFFINITY OF ITS CORE PROMOTER

Vishnevsky O.V.
Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia;
Novosibirsk State University, Novosibirsk, Russia
e-mail: oleg@bionet.nsc.ru

Key words: *TATA-box, TBP, transcription regulation, promoters*

Motivation and Aim: Binding of TATA Binding Protein (TBP) and TATA-box of eukaryotic core promoters is a key stage of preinitiation complex assembly and transcription initiation. An appearance of a huge amount of new precise experimental data about gene expression in different tissues and organs makes it possible to estimate dependence of gene expression on nucleotide context of the promoter to reveal the key features of the gene regulatory regions that determine gene expression.

Methods and Algorithms: An approach for prediction of TBP-affinity and a TATA-box of core promoter region has been suggested [1]. It is based on 4-step model of TBP binding to the TATA-box: nonspecific binding of TBP and DNA, sliding, stoppage, and stabilization. The method was applied to analyze of ~14300 RefSeq mouse promoters within [-90; +1] region relative transcription start site. Dependence of the core promoter nucleotide context on the gene expression in different regions of a mouse brain was estimated. An information about gene expression was obtained from Allen Brain Atlas [2], which contains colorimetric *in situ* hybridization (ISH) data concerning expression of more than $2*10^4$ genes in ~$5*10^4$ voxels (200μm^3 cubes) of a mouse brain.

Results: An analysis of the Allen Brain Atlas data has revealed a reliable correlation ($R = 0.35$, $p < 0.001$) between average gene expression with an amount of the voxels with non-zero gene expression. A positive correlation between predicted TBP affinity with a coefficient of variation of gene expression has been shown ($R = 0.21$, $p < 0.001$).

Conclusion: It has been shown that genes with higher level of average expression are expressed in higher number of the mouse brain regions. And the higher TBP affinity of gene core promoters, the higher expression level variability. The results are consistent with the earlier findings [3] and may be explained by the increased amplitude of the possible expression level of genes with TATA – positive core promoters.

Acknowledgements: The work was supported by the SB RAS project 136.

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.
2. E.S. Lein et al. (2007) Genome-wide atlas of gene expression in the adult mouse brain, *Nature,* **445:** 168-176.
3. Ponomarenko P.M. et al. (2010) A precise equation of equilibrium of four steps of TBP binding with the TATA box for prognosis of phenotypic manifestation of mutations, *Biophysics (Moscow),* **55**: 358-369.

TREATMENT OF CELLS K562/4-NQO AND K562/2-DQO WITH CHEMICAL COMPOUNDS OF MULTIDRUG RESISTANCE LEADS TO APOPTOSIS

Volkova T.O.*, Bagina U.S., Zykina N.S., Malysheva I.E., Poltorak A.N.
Petrozavodsk State University, Petrozavodsk, Russia
e-mail: VolkovaTO@yandex.ru
** Corresponding author*

Key words: *caspases, apoptosis, differentiation, myeloid cancer cell lines*

Motivation and Aim: Many chemical compounds, including anti-cancer drugs, initiate *in vitro* and *in vivo* apoptosis of cancer cells. Nevertheless, clinical applications of cytostatic drugs often lead to a development of the phenotype of multidrug resistance (MDR) in cancer cells. One of the possible mechanisms of the MDR is change in genes and proteins that control apoptosis and cell survival. In the presented work, we studied changes in expression of the genes encoding caspases-3, 6, 9 as well as their enzymatic activity in the cells of K562 line and it's sublines resistant to 4-nitroquinoline-1-oxide (K562/4-NQO) and 2-(4'-dimethylamino styryl)quinoline-1-oxide (K562/2-DQO) upon treatment of the inhibitors of spindle microtubules – kolhicine and nokodazol.

Methods and Algorithms: Gene expression was assessed by real-time PCR, with SYBR Green as a fluorophore. Amplification was performed on iQ5 using kits combined with reverse transcription (iScript One-Step RT-PCR Kit). Caspase activity was measured according to the manufacturer's protocol ("BioRad", USA) with AFC-labeled substrates. Caspase substrates were as follows: 3 – DEVD (Asp-Glu-Val-Asp), 6 – VEID (Val-Glu-Ile-Asp), 9 – LEHD (Leu-Glu-His-Asp). DNA fragmentation (apoptosis) was estimated by electrophoresis in 2% agarose gel.

Results: It is established that the cells K562/2-DQO are resistant to kolhicine and nokodazol (EC_{50} 0,120 mkM and 0,127 mkM correspondingly) when compared with the parental cell K562 (EC_{50} 0,002 mkM и 0,009 mkM, correspondingly) and K562/4-NQO (EC_{50} 0,0004 mkM и 0,0023 mkM, correspondingly). Down regulation of caspase genes is statistically significant as compared to parental K562, which might be one of the reasons of developing resistance to 4-NQO and 2-DQO. Upon treatment of these cells with kolhicine and nokodazol, we observed a statistically significant up-regulation of caspases-3, 6, 9 in the parental line K562 and K562/4-NQO. In the cells of K562/2-DQO we observed activation of only caspase-6. Changes in enzymatic activity of caspases correlated with the levels of mRNA.

Conclusion: 1. We identified different sensitivity of cells K562/4-NQO and K562/2-DQO towards inhibitors of spindle microtubules (kolhicine and nokodazol);

2. Treatment of cells K562/4-NQO with kolhicine and nokodazol resulted in up-regulation of caspases-3, 6, 9 and subsequently to apoptosis. Similar effect in the cells of K562/2-DQO was observed only for caspase-6, and apoptosis was not observed.

This project was supported by grants: № 11.G34.31.0052 and № 1642.2012.4.

ORGANIZATION, EVOLUTION, STRUCTURE AND COMPUTATIONAL PREDICTION OF HUMAN miRNAs

Vorozheykin P.*[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, hidden Markov model, transposable elements, evolution*

Motivation and Aim: miRNAs are a large family of non-coding small RNAs that control gene expression either by the mRNA cleavage or by the translation arrest. Computational methods could become a reliable approach for prediction of tissue-specific or lowly expressed miRNAs and their precursors. While miRNA functioning becomes increasingly better investigated, the organization and evolutionary origins of miRNAs are not well understood.

Methods and Algorithms: In the study we used the sequences of known human mature miRNAs and pre-miRNAs (miRBase, release 17), human genome (ncbi.nlm.nih.gov, release 37.1) and transposable elements from Repbase (release 16.10). The algorithm of the miRNA sites prediction is generalized by using HMM and taking into account the miRNA overlapping. We have modified the algorithms of Nam *et al*. [1] using the novel HMM and the context-structural characteristics of precursors.

Results and Conclusion: The analysis of the genomic distribution of known miRNAs has showed the statistical correlations in their positioning up to 5000 nt, the miRNAs are distributed non-uniformly among the chromosomes while they were randomly copied [2].

We have found that the number of TE-derived pre-miRNAs is growing fast and now is more than 1/5 of all known precursors; endogenous retroviruses are the most probable novel miRNA source. The agreement of the time of large-scale genomic with phenotypic changes may indicate a crucial role of TEs in the primate's evolution and an involvement TE-derived miRNAs into fundamental regulatory processes. DNA-transposon *Made1* has played the major role in the pre-miRNA genomic copying [2, 3].

Our human pre-miRNA prediction shows higher sensitivity than ProMiR [1] with the same specificity. Using the data of loop positioning within miRNA precursors the HMM algorithms of Nam *et al*. for human miRNA prediction [1] were modified to obtain about 2 nt more accurate miRNA boundary.

Acknowledgemen: The work was supported by the RAS Presidium Program #28 "The problem of life origin and biosphere formation".

References:

1. J.-W. Nam, K.R. Shin, J.J. Han, et al. (2005) Human miRNA prediction through probabilistic co-learning model of sequence and structure. *NAR*, **33**(11): 3570-3581.
2. I. I. Titov and P. S. Vorozheykin. (2011) Analysis of miRNA duplication in the human genome and the role of transposon evolution in this process. *Russian Journal of Genetics: Applied Research*, **1**(4): 308-314.
3. J. Piriyapongsa, I King Jordan. (2007) A Family of Human MicroRNA Genes from Miniature Inverted-Repeat Transposable Elements. *PLoS ONE,* 2 :e203.

CONTEXTUAL DNA FEATURES SIGNIFICANT FOR THE DNA DAMAGE BY THE 193 NM ULTRAVIOLET LASER BEAM

Vtyurina N.N.[1, 2], Grokhovsky S.L.[1], Vasiliev A.B.[2], Titov I.I.[3], Ponomarenko P.M.[3], Ponomarenko M.P.*[3], Peltek S.E.[3], Nechipurenko Yu.D.[1, 2], Kolchanov N.A.[3, 4, 5]
[1] Engelhardt Institute of Molecular Biology RAS, Moscow, Russia; [2] Lomonosov Moscow State University, Moscow, Russia;[3] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia; [4] Novosibirsk State University, Russia; [5]National Research Centre "Kurchatov Institute", Moscow, Russia
e-mail: pon@bionet.nsc.ru
** Corresponding author*

Motivation and Aim. Ultraviolet **(**UV-) lasers are used in bioluminescence, cancer therapy, cosmetology, microsurgery, footprinting and many other ways of working with living organisms. A bulk of experiments had shown that laser-induced UV-radiation damages DNA. The most common UV-damage of the DNA at wavelengths ≤ 290 nm is the formation of 7,8-dihydro-8-oxoguanine and other oxidation products of guanine cation (G^+), leading to DNA strand breaks or nucleotide substitutions, which is fundamentally different from the DNA damage caused by other effects (e.g. ultrasound). These more frequent UV-damage of guanine is commonly associated with the energy hν of the absorbed photon that is enough for the escape of an electron (e-) from DNA with the formation of a "hole" (DNA^+). The "hole" then "walks" along the DNA helix and usually stops with the formation of G^+ as a result of a lower ionization threshold of G compared with A, T and C. Than G^+ is immediately attacked by free radical-anions. It was shown that the frequency of UV-damages of different guanines within one nucleotide sequence can vary widely, indicating its dependence on the nucleotide context of the area surrounding each guanine. However, this dependence is still unknown. The question of a systematic *in silico* analysis of the characteristics of the DNA nucleotide context affecting the frequency of guanine UV-damages has not yet been raised.

Methods and Algorithms. We are the first to carry out a systematic in silico analysis of the contextual characteristics of the DNA that affect the frequency of guanines damages at UV-radiation [Vtyurina et al., Biophysics (Mosc) 2011. 56: 410–41].

Results. Characteristics of the local environment of G that we found in this paper and that are significantly affecting the frequency of UV-damaged G, include: 1) consensus ttaaagcHtcg-actgc that is the significantly rare, 2) Position-Weight Matrix, PWM; 3) amount of the tetranucleotide YNVW upstream of G, and 4) estimated frequency of contact with the histone-like protein HU *Escherichia coli*, defects in which increase the frequency of UV-damages. We are the first to construct a linear-additive estimate of the UV-damaged G frequency upon its environment {S(G)} in the DNA strand, such as:
$$\mathbf{f\{S(G)\}=0.69-0.07N_{consensus}\{S(G)\}+0.19PWM\{S(G)\}+0.22YNWV_F\{S(G)\}+0.07P_{Histone\text{-}like}\{S(G)\}.}$$

Conclusion. We are the first to obtain the significant correlation between prediction *in silico* and measurements *in vitro* of the UV-damages frequency in both analyzed [Vtyurina et al., Biophysics (Mosc) 2011. 56: 410–41] ($r=0.679$, $\alpha<10^{-6}$) and independent experiment [Melvin T. et al. NAR 1998. 26:4935-42] ($r=0.821$, $\alpha<0.005$).

Acknowledgments. We thank RFBR 11-04-02001, 11-04-01888, 12-04-01584, НШ-5278.2012.4, and, also, The Russian Academy of Sciences, Presidium Programs "Molecular and Cellular Biology", "Biodiversity", "Biosphere Origin and Evolution".

COMPUTATIONAL NEW SPLICE VARIANTS DISCOVERY USING SINGLE MOLECULE SEQUENCING TECHNOLOGY

Vyatkin Yu.V. [1], Shtokalo D.N.[1, 2, 3], Kapranov P.[3], St. Laurent G.C. III*[3]
[1] *AcademGene LLC, Novosibirsk, Russia;*
[2] *A.P.Ershov Institute of Informatics Systems SB RAS, Novosibirsk, Russia;*
[3] *St.Laurent Institute, Providence, USA*
e-mail: georgest98@yahoo.com
** Corresponding author*

Key words: *alternative splicing, splice junctions discovery, next generation sequencing*

Motivation and Aim: Alternative splicing is an essential post transcriptional regulation phenomenon in eukaryotes by which exons of pre-mRNAs are reconnected in multiple ways. The discovery of alternative splicing events is important for understanding of cell functioning. Recent growth of Next Generation Sequencing technologies [1] provides a huge stream of new data, which leads to new discoveries in genomics [2]. We present a computational method of discovering alternative splice variants based on single molecule sequencing data. The method applied to sequenced *Drosophila melanogaster* transcriptome allowed us to discover previously unknown 435 splice events.

Methods and Algorithms: All possible exon-exon junctions both canonical and putative non-canonical were created combinatorically on per gene basis from all described genes from FlyBase 5.17 database. A set of reads sequenced using Helicos single molecule sequencing technology was aligned to the reference of splice junctions by applying Helisphere indexDPgenomic aligner [3] and then filtered for ambiguity. The non-canonical splice junctions that showed good coverage were annotated by a list of already reported spliced ESTs (http://genome.ucsc.edu) and recent studies [2].

Results and conclusion: Being applied to *Drosophila melanogaster* genome with 21 753 described genes, the method allowed us to build around 250 000 putative non-canonical and known canonical splice junctions. About 240 million informative reads of *Drosophila melanogaster* transcriptome were used to find significant expression of 46 648 splice junctions either canonical or non-canonical. Finally, 769 non-canonical splice junctions from 480 genes with an appropriate coverage were discovered, out of which 435 were not found by us to be reported neither by EST nor by similar study [2]. New sequencing technologies provide an opportunity to find new genomic features even in intensively studied genomes like *Drosophila melanogaster*'s.

References:

1. Thompson JF and Milos PM (2011). The properties and applications of single molecule DNA sequencing. *Genome Biology*, **12**: 217.
2. Graveley BR et al (2011). The developmental transcriptome of Drosophila melanogaster. *Nature*, **471**:473–479.
3. Giladi El. et al (2010). Error tolerant indexing and alignment of short reads with covering template families. *Journal of Computational Biology*, **17**: 1279-1293.

A NEW COMPREHENSIVE CLASSIFICATION OF MAMMALIAN TRANSCRIPTION FACTORS USED FOR NETWORK CONSTRUCTION

Wingender E.*, Haubrock M.J.Li
Department of Bioinformatics, University Medical Center Gųttingen, Gųttingen, Germany
e-mail: ewi@bioinf.med.uni-goettingen.de
** Corresponding author*

Key words: *transcription factor classification, DNA-binding domain, transcriptional network*

Motivation and Aim: The fragmentary available information about DNA-binding specificities of mammalian transcription factors (TFs) renders any reconstructed transcriptional network highly incomplete. We therefore expand the sparse knowledge of DNA-binding behavior of individual TFs to all TFs sharing a DNA-binding domain (DBD) of the same (sub-)family, in order to construct a much more realistic reference network and filter this for individual tissues using known expression patterns.

Methods and Algorithms: About 1600 human TFs and their DBDs were identified in UniProt and TRANSFAC and were grouped into families by multiple alignments and phylogenetic analyses. The transcriptional reference network was constructed by predicting TF-target gene relations from potential TF binding sites (TFBS, identified with the TRANSFAC PWM library) in the corresponding upstream sequences, accepting only high-scoring sites that are highly conserved among 5 mammalian genomes ("seed sites"). This reference network was filtered using TF expression data that were retrieved from UniGene and CGAP, resulting in tissue-specific regulatory networks (TRNs).

Results: Following the principles of on an earlier TF classification and its later extensions (1,2), human TFs were now classified into 9 superclasses, 40 classes, 112 families, many of the latter subdivided into altogether 336 subfamilies, based on the 3D topologies (class levels) and sequence similarities (family levels) of their DBDs. Subfamily and family relations allowed a rational expansion of the eight TRNs reconstructed so far, which show remarkable topological differences (e.g., degree and motif distributions).

Conclusion and Availability: The human TF classification is available at http://www.bioinf.med.uni-goettingen.de/projects/tfclassification/ and, as a more extended draft version, at

http://www.edgar-wingender.de/huTF_classification.html.

References

1. E.Wingender. (1997) Classification scheme of eukaryotic transcription factors, *Molekularnaya Biologiya*, **31:** 584-600.
2. E.Wingender. (2008) The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation, *Brief. Bioinform.*, **9:** 326-332.

PHYLOGENETIC ANALYSIS OF ESX HOMEOBOX PROTEIN OF *BUBALUS BUBALIS*

Brijesh Singh Yadav, Md. Faheem Khan, Ajay Kumar
Indian Veterinary Research Institute, Izatnagar. 243122

Key words: *ESX homeobox, Bubalus bubalis, phylogenetic tree, multiple sequence alignments*

Abstract: The protein sequences of Esx homeobax protein of *Bubalus bubalis* was taken from GenBank Database of NCBI (National Center for Biotechnology Information). Multiple sequence alignments of these proteins for phylogenetic analysis were generated using Clustal W server and the data obtained was analyzed to determine the similarity between the protein sequences of different species. The protein resulted into 569 positions out of which 305 are parsimony informative, 346 are variable sites, and conserved sites were 106 and 32 singleton sites. The most frequent amino acid of these sequences is Alanine, Glutamine, Glycine, Leucin, Proline and Serine. The conserved domain regions search in these proteins from amino acid 220-260, 287-320, 387-400 regions. The amino acid regions from 40-80, 100-190 and 240-320 have lower hydrophobicity, subjected to more conserved part of the protein. Constructed phylogenetic tree of Esx homeobox protein was categorized into two major clades on the basis of their evolutionary distances and bootstrap values calculated through neighbor-joining method. Esx homeobox protein of *Bubalus bubalis* is closely related to homeobox 1 like protein and homeobox protein ESX1 of *Bos tourus* with 100 % boot strap value. These proteins are very much similar with ESX homeobox 1 and protein of *Homo sapience* with 100% boot strap value.

IMPLEMENTING PERMUTATION TEST ON GPU

Yakimenko A.A.*[1], Gunbin K.V. [2], Khaitredinov M.S. [1]
[1] *Institute of Computational Mathematics and Mathematical Geophysics SB RAS, Novosibirsk, Russia;*
[2] *Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia*
e-mail: al--le@yandex.ru
** Corresponding author*

Key words: *permutation test, GPU*

Motivation and Aim: It is well known that statistical inference in genomics and proteomics complicated by the necessity to simultaneous test hundred of statistical hypotheses. To correct the occurrence of false positives various computationally fast analytical multiple testing corrections are frequently used [1]. However, analytical corrections often lead to under- or overestimation of the statistical significance. The permutation test generates null-hypothesis by re-sampling N times total number of observations in a population sample. This approach is more accurate but computationally very intensive. At the same time permutation test can be parallelized easily. Thus, the aim of this wok was the acceleration of the permutation test by its implementation on GPU.

Methods and Algorithms: The parallel version of the program was written in C++. For access to the resources of the graphic accelerator, CUDA (NVIDIA) technology was used [2]. To perform random number generation the Mersenne twister pseudo-random number generator was used [3]. Arrays and associative arrays are widely used in our program; we used standard classes of C++ std::vector and std::map for these implementation.

Results: Comparison of run-time implementation of a coherent program for the CPU and its parallel counterpart, made in a system with a graphics accelerator NVIDIA GTX495, demonstrates speedup by factor of ten. It is important to note that our program used simple and universal format for input and output data. The input of the program is the text tabulation delimited file containing 3 columns: the observation identifier (for example, gene ID), text information about the observation subdivided into terms (for example, GeneOnlology terms), and the numerical value describing the observation (for example, gene expression level or evolutionary conservation). The output is the probabilities of the observation for all terms in the input table.

Conclusion: Permutation test was implemented and tested on the GPU architecture. A significant acceleration of execution time compared to the sequential version was shown.

Acknowledgements: The work supported by SB RAS project 130 and 39, RF State Contracts 07.514.11.4011 and 07.514.11.4003.

Availability: Linux x64-86 binary file available upon request.

References:

1. So H.C., Sham P.C. (2011) Multiple testing and power calculations in genetic association studies. *Cold Spring Harb Protoc.*, doi:10.1101/pdb.top95.
2. CUDA™, http://www.nvidia.com/object/cuda_home_new.html
3. Mersenne Twister for Graphic Processors (MTGP), http://www.math.sci.hiroshima-u.ac.jp/~m-mat/MT/MTGP/index.html

USING PALEOGENOMICS TO STUDY THE ORIGIN AND MECHANISM OF DIVERSIFICATION IN VERTEBRATES: THE CASE OF THE RELAXIN FAMILY PEPTIDES AND THEIR RECEPTORS

Yegorov S., Good S.V.
University of Winnipeg, Canada, e-mail: yegorovsrg@gmail.com

Motivation and Aim: We demonstrate how ancestral genome reconstructions can be used to delineate the origin and duplication history of genes using as example three gene families: relaxin and insulin-like peptides (RLN/INSL) and their two evolutionarily distinct types of receptors (RXFP1/2 and RXFP3/4). *Methods and Algorithms:* Using their exact map positions, we mapped the RLN/INSL, RXFP and INS/IGF genes found in human, medaka and chicken to their corresponding chromosomal segments. These chromosomal segments were then matched to the linkage groups in ancestral genomes [2-5], which allowed us to resolve the positions of the focal genes at consecutive stages of vertebrate genome evolution. Furthermore, we sought support for paleogenomics models-derived data by performing small scale synteny and phylogenetic analyses (in PHYML), for which we used 235 genes (both annotated and not) from 25 vertebrates available in Ensembl (v.60) and NCBI. Additionally, we traced the origin of our focal genes in the pre-2R taxa by searching primitive chordate genomes, including *Ciona*, amphioxus, sea urchin, and fruit fly.

Results (summarize the scientific advance or novel results of the study):

We show that the numerous vertebrate RLN/INSL and RXFP genes are products of an ancestral receptor-ligand system that originally consisted of three genes, two of which apparently trace their origins to invertebrates. Subsequent diversification of the system was driven by whole genome duplications (WGD, 2R and 3R) followed by almost complete retention of the ligand duplicates in most vertebrates, but massive loss of receptor genes in tetrapods. The majority of 3R-duplicates retained in teleosts are potentially involved in neuroendocrine regulation. We also infer that the ancestral RXFP3/4 receptor and RLN/INSL ligand genes may have been syntenically linked in the pre-2R genome, and show how syntenic linkages among ligands and receptors have changed in different lineages.

Conclusion This study ultimately shows the broad utility, with some caveats, of incorporating paleogenomics data into understanding the evolution of gene families. It provides a mechanism to resolve the origin of genes and the relationship of orthology and paralogy in gene families. To the best of our knowledge, we present the first complete reconstruction of the evolutionary relationship of both RLN/INSL peptides and their receptors in vertebrates.

References

1. Yegorov, S. and Good, S. (2012) Using Paleogenomics to Study the Evolution of Gene Families: Origin and Duplication History of the Relaxin Family Hormones and Their Receptors, *Plos One*, **7**, e32923.
2. Nakatani Y, Takeda H, Kohara Y, Morishita S (2007) Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates. Genome Res 17: 1254–1265.
3. Kasahara M, Naruse K, Sasaki S, Nakatani Y, Qu W, et al. (2007) The medaka draft genome and insights into vertebrate genome evolution. Nature 447: 714–719.
4. Kemkemer C, Kohn M, Cooper DN, Froenicke L, Hogel J, et al. (2009) Gene synteny comparisons between different vertebrates provide new insights into breakage and fusion events during mammalian karyotype evolution. BMC Evol Biol 9: 84.
5. Putnam NH, Butts T, Ferrier DE, Furlong RF, Hellsten U, et al. (2008) The amphioxus genome and the evolution of the chordate karyotype. Nature 4537198: 1064–1071.

GTRD: ANNOTATING HUMAN GENOME WITH REGULATORY ELEMENTS USING CHIP-SEQ DATA

Yevshin I.*[1, 2], Kondrakhin Yu.[1, 3], Sharipov R.N.[1, 4], Valeev T.[1, 2], Kolpakov F.A.[1, 3]
[1] *Institute of Systems Biology, Ltd, Novosibirsk, Russia;*
[2] *Novosibirsk State University, Novosibirsk, Russia;*
[3] *Design Technological Institute of Digital Techniques, SB RAS, Novosibirsk, Russia;*
[4] *Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia*
e-mail: ivan@systemsbiology.ru
** Corresponding author*

Key words: *transcription factor binding sites, ChIP-seq, position weight matrix approach*

Motivation and Aim: Regulation of gene transcription is mediated by chromatin modifications and cooperative action of transcription factors (TFs). ChIP-seq technology allows to identify DNA regions bound by individual TFs in the whole genome. Huge amounts of data generated by ChIP-seq experiments should be processed and summarized to produce unambiguous database of regulatory elements in the whole human genome.

Methods and Algorithms: We collected raw ChIP-seq data known from literature, GEO, SRA and ENCODE databases for human and mouse. All the data were processed in the uniform way: sequenced reads were aligned to the reference genome using Bowtie[1] and approximate binding regions of TFs were identified using MACS[2]. We applied our IPS-align method[3] to these regions to construct position weight matrices (PWM) for each TF. PWMs were used for theoretical prediction of TF DNA binding sites. The procedure for learning PWMs from several ChIP-seq experiments automatically selects the best model based on the analysis of receiver operating characteristic. We use hierarchical classification of TFs by their DNA-binding domain (http://www.edgar-wingender.de/huTF_classification.html) to make generalized PWMs for TF classes in the case when DNA binding sites of similar TFs cannot be distinguished by their sequences.

For automatic database updates special workflows for BioUML platform (http://www.biouml.org) were developed.

Results: We created the Gene Transcription Regulation Database (GTRD) database using 1115 ChIP-seq experiments for 128 TFs. Genomic regions identified for several cell lines and conditions coupled with genome-wide theoretical prediction allow unambiguously annotating the whole human genome with the set of DNA regulatory elements.

Availability: http://biouml.org/gtrd.shtml

Acknowledgements This work was supported by the RFBR (grant Nr. 10-04-01524-a) and the Presidium of the RAS (The "Basic Science for Medicine" Program, grant Nr.33).

References:

1. B. Langmead et al. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome, Genome Biol, 10:R25
2. Zhang et al. (2008) Model-based Analysis of ChIP-Seq (MACS), Genome Biol, 9(9): R137.
3. I. Yevshin et al. (2012) IPSscan: the extended matrix method for prediction of transcription factor binding sites. Proceedings of BGRS'2012. Submitted.

IPSSCAN: THE EXTENDED MATRIX METHOD FOR PREDICTION OF TRANSCRIPTION FACTOR BINDING SITES

Yevshin I.*[1, 2], Kondrakhin Yu.[1, 3], Sharipov R.N.[1, 4]
[1] *Institute of Systems Biology, Ltd, Novosibirsk, Russia;*
[2] *Novosibirsk State University, Novosibirsk, Russia;*
[3] *Design Technological Institute of Digital Techniques, SB RAS, Novosibirsk, Russia;*
[4] *Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia*
e-mail: ivan@systemsbiology.ru
** Corresponding author*

Key words: *transcription factor binding sites, prediction, position weight matrix approach*

Motivation and Aim: Since its introduction in 2007, ChIP-Seq has become the most powerful experimental technique for the genome-wide study of interactions between transcription factors (TFs) and DNA. As a rule, a single ChIP-Seq experiment generates millions of short reads, and TF-binding regions (TF-BRs) are identified by applying peak detection algorithms (PDAs) to the set of reads aligned to the reference genome. However, comparative analysis of nine PDAs [1] demonstrated that biological conclusions could change dramatically when the same raw ChIP-Seq dataset was processed using different algorithms. Furthermore, the application of theoretical methods for prediction of binding sites (BSs) in addition to PDAs allowed to increase the accuracy of BSs identification [2]. Thus, despite the emergence of the ChIP-Seq technology, development of the advanced methods for theoretical identification of TF-BSs is still actual.

Methods and Algorithms: A novel method, IPSscan, has been developed. It represents the extension of common position weight matrix (PWM) approach. The IPSscan exploits individual probability scores (IPSs) that were derived analytically under an assumption of independence of site positions. IPSs depend on the common additive scores and the nucleotide contents of the both flanks of the site cores.

Results: We demonstrated that application of the proposed IPSscan leads to essential increase of accuracy of TF-BSs prediction, on average, by 2-3-fold. We also have found that the most reliable TF-BRs were the mostly enriched with TF-BSs predicted by IPSscan.

Availability: The predicted TF-BSs are available in the GTRD database (http://biouml.org/gtrd.shtml) [3] developed on the basis of the BioUML platform (http://biouml.org/)

Acknowledgements This work was supported by the RFBR (grant Nr. 10-04-01524-a) and the Presidium of the RAS (The "Basic Science for Medicine" Program, grant Nr.33).

References:

1. T.D. Laajala *et al*. (2009) A practical comparison of methods for detecting transcription factor binding sites in ChIP-seq experiments, *BMC Genomics*, **10:**618.
2. O. Wallerman O. *et al*. (2009) Molecular interactions between HNF4a, FOXA2 and GABP identified at regulatory DNA elements through ChIP-sequencing, *Nucleic Acids Res.*, 37: 7498-7508.
3. I. Yevshin *et al.* (2012) Gene Transcription Regulation Database: annotating human genome with regulatory elements using ChIP-seq data, *Proceedings of BGRS'2012. Submitted.*

THE p53IPS MODEL FOR SELF-ORGANIZING META-PREDICTION OF p53-BINDING SITES AND p53-TARGET GENES

Yevshin I.*[1, 2], Kondrakhin Yu.[1, 3], Sharipov R.N.[1, 4]
[1] *Institute of Systems Biology, Ltd, Novosibirsk, Russia;*
[2] *Novosibirsk State University, Novosibirsk, Russia;*
[3] *Design Technological Institute of Digital Techniques, SB RAS, Novosibirsk, Russia;*
[4] *Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia*
e-mail: ivan@systemsbiology.ru
** Corresponding author*

Key words: *p53-binding sites, p53-target genes, meta-prediction, position weight matrix approach*

Motivation and Aim: Transcription factor p53 has been studying intensively because it plays essential role in many key biological processes such as control of DNA integrity, apoptosis, opposition to cancer development and the complex scheme of intracellular interactions mediating both activating and suppressing signals. Therefore prediction of p53 binding sites is actual so far. Nonetheless, the existing algorithms are far from prediction of the complete set of p53-binding sites. The main aim of this work was to narrow the gap in identification of p53-binding sites by applying sophisticated prediction methods.

Methods and Algorithms: The proposed p53IPS model consists of four self-organizing levels, and sites predicted by previous levels are used as training sets for further levels. The p53IPS incorporates different aspects and information on the functioning of p53: the various sets of p53-binding regions identified by ChIP-Seq and ChIP-Chip experiments, specific histone modification pattern, p53-responsive genes identified by microarray experiments, binding regions of p53 coregulators AP1 and CBP and p53-binding motifs predicted theoretically by our prediction method. In turn, p53IPS model takes advantage of optimal matrix and individual probability scores (IPSs) instead of additive scores used in the common position weight matrix approach. The transition from the common approach to IPS leads to improvement of prediction accuracy.

Results: Initially, we constructed the optimal position weight matrix for prediction of p53-binding sites. Then we created a self-organized model, p53IPS. Finally, about half of the predicted sites were filtered out because were classified as associated with SINE/Alu or LINE/L1 repeats. In the frame of this model we have meta-predicted 1498 reliable p53-binding sites that were mapped to regulatory regions of 1292 genes. Only about 5% of them have been known as genes regulated directly by p53. According to Gene Ontology, significant part of 1292 genes underlies functions of the central neural system and the heart, participates in angiogenesis, lipid metabolism and inflammation. This knowledge can be potentially used in drug development and treatment of a range of human diseases (including metabolic).

Conclusion: The p53IPS model has been developed giving advantages in comparison to the existing methods of the p53 site prediction. Using this model about 1200 genes have been recognized as new direct targets of p53 and this fact will be validated in experiments.

Acknowledgements The work was supported by the RFBR grant Nr. 10-04-01524-a and the grant Nr.33 of the "Basic Science for Medicine" Program of the Presidium of the RAS.

IDENTIFICATION OF NEW METHYLATION-REGULATED GENES AS MOLECULAR TARGETS FOR PHARMACEUTICAL INTERVENTION AND DIAGNOSIS BASED ON NOTI MICROARRAYS

Zabarovsky E.R.
Author affiliations: Karolinska Institute and Linkoping University, Stockholm, Sweden

Key words: *NotI microarrays, epigenetics, biomarkers, early diagnosis, prognosis*

Motivation and aim: To find biomarkers for early diagnosis, prediction and prognosis of lung, prostate, ovarian, cervical and kidney cancer.

Methods: The project is based on large-scale sequencing of human NotI clones performed by us. We showed that a human cell contains up to 10.000 unmethylated NotI sites each of which is associated with a gene(s). We recently developed a new type of microarrays– NotI microarrays. This new techniques allows high-throughput genome scanning for both genetic and epigenetic (methylation) changes. Methylation was found to be a basic and probably the most important mechanism regulating gene expression. Capability to establish differences between normal and pathological cells will be instrumental for cloning disease-causing genes and prevention cancer.

Results: We analyzed more than 450 different cancer samples (incl. lung, cervical, breast, ovarian) and found new tumor suppressors and oncogenes. More than 50 genes were found to be methylated in these tumors in more than 25% of cases. Among these genes only a few genes were already known TSGs. Majority of the found genes were previously not shown to be involved in carcinogenesis, like MINA, TRH, PPP2R3A, FLJ44898, FGD5.

Methylation of 32 genes was confirmed using bisulfite sequencing in more than 100 tumor samples. Downregulation of 20 selected genes was also observed in these cancers. Biomarkers for early diagnosis and prognosis were developed for ovarian, cervical, lung, renal and prostate cancers.

Conclusion: Using obtained data a set of 18 markers (BHLHB2, FBLN2, FLJ44898 (EPHB1), GATA2, GORASP1, Hmm210782 (PRICKLE2), Hmm61490, ITGA9, LOC285205, LRRC3B, MINA, MITF, MRPS17P3, NKIRAS1, PLCL2, TRH, UBE2E2, WNT7A) that allow to discriminate/diagnose SCC and ADC, two main types of NSCLC (early detection, progression, metastases) with probability more than 95% was created and is under validation now. Similar sets were created for diagnosis/prediction of prostate,ovarian, cervical and kidney cancers. After validation, chips with these genes will be checked in clinical tests.

Acknowledgements: Kashuba Vladimir I, Grigorieva Elvira V., Haraldson Klas, Dmitriev Alexey, A, Pavlova Tatiana, Krasnov Georgy, Braga Eleonora.A., Yenamandra Surya P, Lerman Michael I.

HOW LONG SEQUENCED GENOME CAN REMAIN STABLE

Zakharenko L.P.*[1, 2], Bak T.P. [1], Ignatenko O.M. [2]
[1] *Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia;*
[2] *Novosibirsk State University, Novosibirsk, Russia*
e-mail: zakharlp@bionet.nsc.ru
** Corresponding author*

Key words: *transposable elements, sequenced genomes, genome instability*

Motivation and Aim. Many genomes are sequenced. But how long is it possible to consider they stable? Especially it concerns stability of an arrangement in a genome of transposable elements (TEs) owing to their high mobility. *Methods and Algorithms.* We analyzed by FISH distribution of some *Drosophila melanogaster* TEs 20 years later after receiving library of clones which were used for sequences. *Results.* We have shown that rate of TEs moving varies from 10^{-2} to 10^{-4} in a genome of the isogenic *y cn bw sp* strain. If total number of TEs makes about 1500 copies on a Drosophila genome, and average speed of TEs moving is 10^{-3}, in each generation at least 1.5 TE change the position. In 20 years passed about 200 generations, that is about 300 Drosophila TEs or the fifth part of TEs changed there's positions. Strangely enough, rate of moving of TEs doesn't correlate with level of their expression and with number of full-size copies. So transposable element *blood* is presented in a genome by two tens full-size copies with an open reading frames and according literature *blood* is expressed in generative tissue. Nevertheless *blood* is extremely stable in this line. At the same time *hobo*-element is extremely movable in *y cn bw sp* genome though only one full-size copy and two tens of defective *hobo* is found in it. Moreover full-size *hobo* transcript wasn't found in any experiment. Discrepancy of level of a transcription of mobile elements, and speeds of their introduction in a genome observe also in other works (1, 2). According our data recombination between different sorts of repeats has significant influence on the TEs distribution in Drosophila genome. *Conclusion.* Thus, first, it is impossible to substitute speed of moving of TEs for level of its expression. Secondly, the genome is much more unstable, than it is accepted to think. However, despite high mobility of mobile elements a phenotype and genetic characteristics of the line under analysis didn't change. That is the importance of moving of mobile elements in life of population is probably strongly exaggerated. The instability of a genome caused by TEs moving, changes genome structure, but affects a little or doesn't affect mainly at all a phenotype. *Availability*: http://flybase.org/static_pages/lists/dmel_te.html. *Acknowledgements*: This work was partially funded by program of Russian Academy of Sciences "Biodiversity" (grant B 27-29) and program "Wildlife" (grant 30-30).

References:

1. Spradling et al.(2011) Drosophila *P* elements preferentially transpose to replication origins. PNAS 108 :15948-15953
2. Esnault et al.(2011) Intrinsic characteristics of neighboring DNA modulate transposable element activity in *Drosophila melanogaster.* Genetics. 187:319-331.

DYNAMIC MODEL OF ANAEROBIC ENERGY METABOLISM OF YEAST *SACCHAROMYCES CEREVISIAE*

Zakhartsev M.*[1], Lapin A. [2], Reuss M. [2]
[1] *IPMB, University of Heidelberg, Germany;*
[2] *Center Systems Biology (CSB), University of Stuttgart, Germany*
e-mail: zakhartsev@uni-heidelberg.de
** Corresponding author*

Key words: *AXP-paradox, energy metabolism, dynamic metabolic model*

Motivation and Aim: Cellular energy metabolism, besides conversion of energy flow for metabolic purposes, is a metabolic hub that interfaces metabolic modules through which a metabolic perturbation can propagate from one module to another, thus implementing a signaling role. At steady state conditions, an adenylate pool (AXP = ATP + ADP + AMP) usually is assumed operating under 'conserved moiety' mass-law. However, fast metabolic perturbation experiments (*e.g.* glucose-pulse) have revealed inconsistency of this assumption on a minute timescale, where AXP pool operates as an 'opened pool'. As a result of the glucose-pulse the entire AXP pool shrinks in a matter of 30 seconds and replenishes only in 5 minutes (so-called 'ATP-paradox'). This phenomenon was observed long time ago however only recently a molecular mechanism underlying this regulation was elucidated. In course of a transition induced by substrate perturbation the excess of AMP is ousting into inosine and hypoxanthine via salvage reactions and then adenylate pool is replenished through both *de novo* and salvage pathways. Consequently the question has arisen: what metabolic meaning does the ATP-paradox have for cellular homostasis?

Methods and Algorithms: To better understand this phenomenon we have performed glucose pulse experiment on anaerobically growing yeast in glucose limited chemostat. The transient concentrations of extra- and intracellular metabolites were measured as a function of time after the perturbation. ODE-based MATLAB model of major metabolic regulatory events in glycolysis, pentose-phosphate pathway, purine *de novo* synthesis, nucleotide salvage reactions, redox balance and biomass growth was developed. The model consists of 43 state variables interconnected by 41 reactions and 5 transport steps. Transient metabolite concentrations were used for parameterization of the dynamic model.

Results: The distinctive feature of this model is that it explains AXP dynamics as 'opened moiety' through purine *de novo* synthesis pathway, purine salvage reactions and biomass growth. The model explains dynamic behavior of all measured metabolites and predicts that the rate of purine *de novo* synthesis shortly increases right after the glucose pulse, which would result in peaking of all intermediates along linear purine *de novo* pathway. This event is one of coupling points between metabolic and genetic regulations. To our knowledge, transient increase of (S)AICAR intermediate from purine *de novo* synthesis pathway can explain earlier observation in whole genome expression profile after the glucose pulse through de-repression of *Bas1* and *Pho2* transcriptional factors, which coordinate upregulation of the purine biosynthesis, sulfur and phosphorus assimilation, methionine and adenine salvage pathways.

Conclusion: Thus, stimulus-response methodology aided with mathematical modeling has allowed us better understand of functional meaning of 'ATP-paradox' as a fast metabolic signal to adapt cell transition from substrate limited to unlimited growth by means of genetic regulation.

NEW CANDIDATE GENES FOR SCHIZOPHRENIA DISORDER

Zakharyan R.V.*, Boyajyan A.S.
Institute of Molecular Biology NAS RA, Yerevan, Armenia
e-mail: r_zakharyan@mb.sci.am
** Corresponding author*

Key words: *schizophrenia, genetic polymorphisms, proinflammatory and chemotactic cytokines, genotyping, polymerase chain reaction with sequence-specific primers*

Motivation and Aim: Schizophrenia is a polygenic and multifactorial disease with strong involvement of the inflammatory component in etiopathogenic mechanisms [1, 2]. However, only limited studies have explored the association of polymorphisms in genes encoding inflammatory mediators with schizophrenia. The present study aimed to investigate the possible association of functional genetic polymorphisms of proinflammatory and chemotactic cytokines including interleukin (IL)-6, tumor necrosis factor-α (TNF-α), monocyte chemoattractant protein 1 (MCP-1), and IL-8 with schizophrenia.

Methods and Algorithms: Genomic DNA samples were isolated from fresh blood of schizophrenia-affected (n=225) and healthy subjects (n=225) according to the standard phenol–chloroform method. All DNA samples were genotyped for targeted single nucleotide polymorphisms (SNP) by polymerase chain reaction with sequence-specific primers (PCR-SSP). All primers for the PCR-SSP were designed using the genomic sequences in the "GenBank" database. The genotypes were assessed for the presence/absence of PCR amplicons specific to the particular alleles using a standard 2% agarose gel electrophoresis followed by ethidium-bromide staining. Distributions of genotypes for the studied polymorphisms were checked for correspondence to the Hardy–Weinberg equilibrium. In order to find potential relevance of targeted SNPs to schizophrenia, the allele and phenotype frequencies in patients and control subjects were compared. The significance of differences between allelic and phenotype frequencies was determined using Pearson's Chi-square test. The odds ratio (OR), 95% confidence interval (CI), statistical power, and Pearson's p-value were calculated in each case.

Results: According to the data obtained, the *IL-6* -174G/C, *TNF-α* -308G/A, *MCP-1* -2518A/G, and *IL-8* +293G/T polymorphisms were significantly associated with schizophrenia.

Conclusion and Availability: The *IL-6* -174G/C, *TNF-α* -308G/A, *MCP-1* -2518A/G, and *IL-8* +293G/T minor alleles seems to be a risk factors for the development of schizophrenia.

References:

1. M. Baron. (2001) Genetics of schizophrenia and the new millennium: progress and pitfalls. Am. J. Hum. Genet., 68: 299-312.
2. A. Monji, T. Kato, S. Kanba. (2009) Cytokines and schizophrenia: microglia hypothesis of schizophrenia. Psychiatry Clin. Neurosci., 63: 257–265.

EVOLUTION OF MITOCHONDRIAL tRNAs

Zaytseva N.A.[1], Kondrashov F.A.*[2], Vlasov P.K. [2]
[1] *Siberian Federal University, Krasnoyarsk, Russia;*
[2] *Centre for Genomic Regulation, Barcelona, Spain*
e-mail: nata.zajtseva@gmail.com
** Corresponding author*

Key words: *tRNA, evolution, consensus sequence, distance*

Motivation and Aim: Structural evolution in tRNA molecules is known [1]. However, the global tRNA evolution within the total sequence space has not been explored. A new computational approach to study the divergence of tRNA sequences is described. The analysis is based on the model of the ongoing expansion of the protein universe [2]. This procedure makes it possible to reconstruct evolution of tRNA of different amino acids. Furthermore the investigation may be of value for solving problems of early translation and the genetic code.

Methods and Algorithms: We have investigated evolution of tRNAs using sequence divergence data. 276 mitochondrial genomes of *Tetrapoda* for each amino acid have been studied. The consensus sequences for each group of tRNA were determined. We calculated the distances from all the sequences to each consensus. We were interested only in sequences, which are closer to consensus sequence of another group, than to their own. Moreover we divided all sequences into four databases: *Mammalia, Aves, Amphibia* and *Reptilia*. These sequences were processed with the same technique.

Results: We found, that 141 histidine, 212 isoleucine and 188 leucine sequences are located closer to consensus of tyrosine than to their consensus sequence. Besides the reverse situation is observed. Significant part of tyrosine sequences are placed nearly histidine, isoleucine and leucine consensus. It means that for some reason tRNAs of these amino acids form cluster in the sequence space. In the center of cluster the tyrosine group is located. It should be noted, that intersections are not found in the small databases (*Mammalia, Aves, Amphibia* and *Reptilia)*. The results agree with theoretical predictions.

Conclusion: The first picture of groups tRNA distribution has been obtained. In our opinion the result is associated with evolutionary events. We need additional research to understand and interpret the findings. Our dataset will be expanded. According to the model when the numbers of sequences increases with increasing, the numbers of intersections will rise. So distribution of tRNA groups varies with classes of organisms, which are in database. Time of the origin of the classes is known; therefore we have an opportunity to observe changes of the allocation of tRNA groups in the time. When the investigation is finished, we will have a better understanding of the nature of tRNA distribution in sequence space. In addition the results have the potential to find new evolutionary principles and patterns.

References:

1. F. Sun, G. Caetano-Anolle (2008) The Origin and Evolution of tRNA Inferred from Phylogenetic Analysis of Structure, *Mol. Evol*., **66(1)**:21-35
2. I. S. Povolotskaya, F. A. Kondrashov (2010) Sequence space and the ongoing expansion of the protein universe, *Nature,* **465**: 922-927.

SYSTEM ANALYSIS OF HUMAN CELL LINE: TRANSCRIPTOME, PROTEOME

Zgoda V.G.*, Tikhonova O., Novikova S., Kurbatov L., Kopylov A., Moskalyova N., Archakov A.I.
Orekhovich Institute of Biomedical Chemistry, Russian Academy of Medical Sciences
e-mail: vic@ibmh.msk.su
** Corresponding author*

Key words: *human leukemia 60 cell line (HL60), retinoic acid, transcriptome, microarray, proteome, mass-spectrometry*

Motivation and Aim: A differentiation of HL60 human promyelocyte leukemia cell line after addition of retinoic acid was used as a model for systems pathway analysis of transcriptome and proteome expression data.

Methods and Algorithms: HL60 transcriptome in several endpoints of cell differentiation was estimated using full human genome microarray platform by Agilent. Proteomes were analyzed using LC-MS/MS platform (LTQ-Orbitrap, Thermo). The results were analyzed to find potential node regulators, e.g., transcription factors responsible for cell differentiation with GeneXplain software platform (www.genexplain.com).

Results: Full genome transcriptomics showed significant differences in expression of more than 1500 genes after treatment of HL60 cell line by retinoic acid. Of them, after differentiation launch, 134, 207, 364, 393 and 1197 genes changed their expression at least 2-fold upon 30 min, 60 min, 3 h, 24 h and 96 h, respectively.

By proteomics with LC-MS/MS, about 1370 proteins were identified in HL60 cell line. Label-free quantitation allows detection of differential expression for 65, 103 and 331 proteins upon 3, 24 and 96 h after retinoic acid addition, respectively.

Data from transcriptomics and proteomics were united and considered using GeneXplain to find pathways and transcription factors responsible for HL60 differentiation under retinoic acid.

Conclusion: GeneXplain platform allows at least general explanation of cell processes during HL60 differentiation using transcriptome and proteome data.

Acknowledgement: The work was funded by Russian Academy of Medical Sciences.

SYNTHETIC LETHALITY WITHIN ONE PATHWAY AND CANCER TREATMENT

Zinovyev A.Yu.*[1, 2, 3], Kuperstein I. [1, 2, 3], Barillot E. [1, 2, 3] , Heyer W.-D. [4]
[1] *Institut Curie, 26 rue d'Ulm, F-75248 Paris France;*
[2] *INSERM, U900, Paris, F-75248 France;*
[3] *Mines ParisTech, Fontainebleau, F-77300 France;*
[4] *University of California, Davis, One Shields Avenue, Davis, CA 95616-8665*
e-mail: andrei.zinovyev@curie.fr
** Corresponding author*

Key words: *synthetic lethality, pathways, mathematical modeling, DNA repair*

Motivation and Aim: A synthetic lethal interaction is usually stated when defects in two non-essential genes cause cell death. Synthetic lethality and synthetic dosage lethality studies in model organisms and human cells give hope to develop cancer drugs that would kill cancer cells very selectively: if a cancer cell has a characteristic deletion or amplification of a gene, then inhibiting or overexpressing another nonessential gene forming a synthetic interaction pair, will lead to specific lethality of cancer cells. For example, some breast cancers are characterized by loss-of-function mutation in BRCA1 gene involved in DNA repair. The PARP1 gene forms a synthetic lethal pair with BRCA1 in cellular models and, therefore, inhibitors of PARP1 for treating BRCA1-deficient breast cancer were developed and went to clinical trials.

The genes from a synthetic interaction pair are generally assumed functioning in two parallel and mutually compensatory pathways (multi-pathway Synthetic Lethality). However, several examples of synthetic lethal relationships involving genes implicated in the homologous recombination DNA repair pathway extend this paradigm. In this situation defects in two genes which function in the same pathway lead to cell death (single-pathway Synthetic Lethality).

Results: We explored the inherent system properties of single-pathway Synthetic Lethality using mathematical modeling. We found that three circumstances are pre-requisites for the single-pathway Synthetic Lethality scenario: reversibility of pathway steps, presence of a compensatory pathway and toxicity of at least one pathway intermediate. Further modeling revealed the potential contribution of synthetic dosage lethal interactions in such a genetic system.

Conclusion: Single-pathway Synthetic Lethality represents a phenomenon which is generally overlooked in interpretation of genetic interactions; however, it can play an important role in explaining many of them. Single-pathway Synthetic Lethality is not limited to DNA repair system but can be generic for many biological pathways.

We discuss implications of single-pathway synthetic lethality to cancer treatment modalities acting through DNA damage.

Availability: The corresponding MATLAB code for the modeling is available from the authors by request.

ACTIVATION OF *CLV3* GENE EXPRESSION IN MODEL OF THE STEM CELL NICHE STRUCTURE REGULATION IN THE SHOOT APICAL MERISTEM

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: *shoot apical meristem, stem cell niche, CLV3, mathematical model*

Motivation and Aim: In spite of many years searching of the mechanism of stem cell niche maintenance in shoot apical meristem (SAM) its particular details are not clear enough. There is a qualitative hypothesis of the interplay between *CLV* and *WUS* genes which are believed to be able to regulate the SAM spatial compartmentalization into central zone (CZ – stem cells) and organizing center (OC). The following is key moment of the hypothesis: *CLV3* expression occurs in the CZ-cells of 3 upper layers, while *WUS* expression occurs in the OC-cells just below CZ; and CLV3 by means of binding with putative receptor CLV1/CLV2 inhibits *WUS* expression, while WUS activates *CLV3* expression. Several models have been proposed that ascribe roles of these genes. Principle feature of our model [1] is interaction between the SAM zones represented as sequence of the activating influences WUS=>Y=>CLV3, where *Y* is a hypothetical gen expressing in the uppermost cells of the SAM. Taking into account the recent experimental studies [2] showed that the WUS protein, after being synthesized in cells of the OC, migrates into the CZ, where it activates CLV3 transcription by binding to its promoter elements, in the present work we study spacial features of activation of CLV3 both by WUS and Y.

Methods and Algorithms: The regulatory network of our previous mathematical model [3] was updated with additional activation arrow (WUS directly activates CLV3). The model was realized in the Cellzilla package (http://cellzilla.info/), and numerically solved to obtain a stationary solution on a 2D domain representing longitudinal section of the SAM.

Results: Different spatial distributions of CLV3 expression domains with different values of WUS- and Y-activating parameters were obtained in frame of the model.

Conclusion: To obtain observed geometry of the CZ when *WUS* directly activates *CLV3* in our model as well as in other models [2, and reffs. therein] it is necessary to propose that this geometry is mainly controlled by a signal coming from a few cells of the SAM tip.

Acknowledgements: The work was partly supported by RFBR grant 11-04-01748-a «Computer analysis and modeling of shoot apical meristem development».

References:

1. S.V. Nikolaev et. al. (2007) A Model Study of the Role of Proteins CLV1, CLV2, CLV3, and WUS in Regulation of the Structure of the Shoot Apical Meristem, *Rus J of Developmental Biology*, **38:** 383–388.
2. R.K. Yadav et. al. (2011) WUSCHEL protein movement mediates stem cell homeostasis in the Arabidopsis shoot apex, *Genes Dev,* **25:** 2025-2030.
3. S.V. Nikolaev et. al. (2010) A reaction-diffusion model of shoot apical meristem compartmentalization based on CLV/WUS interplay *Proc. of the 3rd ICMBB, Pushchino, Russia.*

ACTIVATION OF *CLV3* GENE EXPRESSION IN MODEL OF THE STEM CELL NICHE STRUCTURE REGULATION IN THE SHOOT APICAL MERISTEM

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: *shoot apical meristem, stem cell niche, CLV3, mathematical model*

Motivation and Aim: In spite of many years searching of the mechanism of stem cell niche maintenance in shoot apical meristem (SAM) its particular details are not clear enough. There is a qualitative hypothesis of the interplay between *CLV* and *WUS* genes which are believed to be able to regulate the SAM spatial compartmentalization into central zone (CZ – stem cells) and organizing center (OC). The following is key moment of the hypothesis: *CLV3* expression occurs in the CZ-cells of 3 upper layers, while *WUS* expression occurs in the OC-cells just below CZ; and CLV3 by means of binding with putative receptor CLV1/CLV2 inhibits *WUS* expression, while WUS activates *CLV3* expression. Several models have been proposed that ascribe roles of these genes. Principle feature of our model [1] is interaction between the SAM zones represented as sequence of the activating influences WUS=>Y=>CLV3, where *Y* is a hypothetical gen expressing in the uppermost cells of the SAM. Taking into account the recent experimental studies [2] showed that the WUS protein, after being synthesized in cells of the OC, migrates into the CZ, where it activates CLV3 transcription by binding to its promoter elements, in the present work we study spacial features of activation of CLV3 both by WUS and Y.

Methods and Algorithms: The regulatory network of our previous mathematical model [3] was updated with additional activation arrow (WUS directly activates CLV3). The model was realized in the Cellzilla package (http://cellzilla.info/), and numerically solved to obtain a stationary solution on a 2D domain representing longitudinal section of the SAM.

Results: Different spatial distributions of CLV3 expression domains with different values of WUS- and Y-activating parameters were obtained in frame of the model.

Conclusion: To obtain observed geometry of the CZ when *WUS* directly activates *CLV3* in our model as well as in other models [2, and reffs. therein] it is necessary to propose that this geometry is mainly controlled by a signal coming from a few cells of the SAM tip.

Acknowledgements: The work was partly supported by RFBR grant 11-04-01748-a «Computer analysis and modeling of shoot apical meristem development».

References:

1. S.V. Nikolaev et. al. (2007) A Model Study of the Role of Proteins CLV1, CLV2, CLV3, and WUS in Regulation of the Structure of the Shoot Apical Meristem, *Rus J of Developmental Biology*, **38:** 383–388.
2. R.K. Yadav et. al. (2011) WUSCHEL protein movement mediates stem cell homeostasis in the Arabidopsis shoot apex, *Genes Dev,* **25:** 2025-2030.
3. S.V. Nikolaev et. al. (2010) A reaction-diffusion model of shoot apical meristem compartmentalization based on CLV/WUS interplay *Proc. of the 3rd ICMBB, Pushchino, Russia.*

LARGE SCALE METAGENOMIC CLUSTERING

Zola J.
Department. of Electrical and Computer Engineering,
Iowa State University,
Ames, 50010 Iowa, USA
e-mail: zola@iastate.edu

Key words: *metagenomics, taxonomic clustering, high performance computing, map-reduce, data synopsis*

Motivation and Aim: Metagenomics is the study of a population of organisms by fragmenting and sequencing their collective DNA. It is typically applied to communities of microbial organisms sampled from their native environments where species-wise separation is difficult, expensive, or downright impossible. Taxonomic clustering of species is an important and frequently arising problem in metagenomics. High-throughput next generation sequencing is enabling the creation of large metagenomic samples, while at the same time making the problem of metagenomic data analysis harder due to the short sequence length supported and sampling of hitherto unknown species.

Methods and Algorithms: We present a parallel algorithm for taxonomic clustering of large metagenomic samples. We develop sketching techniques akin to those created for web document clustering to deduce significant similarities between pairs of sequences without resorting to expensive all vs. all comparison. We propose a graph-theoretic formulation of the taxonomic classification problem that supports different taxonomic levels as prescribed by different similarity thresholds. We cast execution of the underlying algorithmic steps as applications of the map-reduce framework to achieve a cloud ready implementation. Apart from solving an important problem in metagenomics, this work demonstrates the applicability of map-reduce paradigm in relatively complicated algorithmic settings.

Conclusion: Our parallel method is highly scalable and can be executed on large map-reduce clusters. Thanks to application of the data synopsis techniques it can be applied to process some of the largest existing metagenomics samples without sacrificing precision.

IDENTIFICATION OF miRNAS OF THREE OPISTHORCHID LIVER FLUKES

Katokhin A.V.*, Afonnikov D.A., Ovchinnikov V.Yu., Vasiliev G.V., Kashina E.V., Mordvinov V.A.
Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
e-mail: katokhin@bionet.nsc.ru
**Corresponding author*

Key words: *miRNAs, liver flukes, Opisthorchiidae*

Motivation and Aim: The small non-coding RNA (mi*RNA* and si*RNA*) are shown to be a significant part of transcriptome of any multicellular organism. They play an important role in regulatory events to determine expression of many protein-coding genes during ontogenesis, morphogenesis and organs functioning. To date the miRNAs are described for 7 species belonging to flat worms. These are three parasitic species of Trematoda class (*Schistosoma mansoni, Schistosoma japonicum, Clonorchis sinensis*); two parasitic species of Cestoda class (*Echinococcus granulosus, Echinococcus multilocularis*) and two species of non-parasitic planarians of Turbellaria class (*Schmidtea mediterranea, Macrostomum lignanum*). The interest for parasitic flat worms miRNAs is generated by the investigations aimed at uncovering intrinsic molecular mechanisms of their developments and pathogenesis in order to design novel anthelmintic tools.

Methods and Algorithms: To experimentally identify the miRNAs in three species of Opisthorchiidae family we extracted total RNAs from 20 adult worms of *Opisthorchis felineus*, 20 - *Opisthorchis viverrini* and 14 - *Clonorchis sinensis*. The *O. felineus* adult worms were manually dissected to produce two separate samples: the distal parts of uterus filled by encapsulated embryos ("eggs") and remaining bodies. The RNAs was also prepared from 5000 *O. felineus* metacercariae (larval stage). To enrich the small RNAs fraction we used the polyethylenglycol precipitation methods by Wang X-W. et al., 2010. After the RNA samples processing with kit "Solid Small RNA Expression kit" the 15 libraries for SOLiD sequencing (Applied Biosystems) were generated.

Results: Upon the sequencing the reads were processed by program suite to filter out the fragments of mRNAs, tRNAs, rRNAs and so on and to produce data suitable for miRNAs identification. The main procedure consisted in mapping of candidate miRNAs on *C. sinensis* genomic contigs available after publication Wang X. et al., 2011. We selected the candidate miRNAs which satisfied a structural requirements common for known miRNAs genes of animals: the 70-100 nt of genome regions around the candidate miRNAs should form a specific stem-loop structures of pre-miRNA. For *O. felineus* the 347 candidate miRNAs were indentified, for *O. viverrini* – 227, for *C. sinensis* – 235. The search in MirBase v.18 allowed us then to determine candidate miRNAs conserved for Platyhelminthes and other Metazoa. Some portion of candidate miRNAs remained without known orthologs so forming a candidate miRNAs specific for either the three opisthorchiids or even for each species under investigation. The candidate miRNAs identified were additionally tested by mapping on *Schistosoma japonicum* genomic contigs.

The comparative analysis of the candidate miRNAs structures and counts from the species-specific and stage-specific libraries was performed and the results were discussed.

Acknowledgements The work was supported by RFBR grant 09-04-12209.

A CENTRAL REGULATORY CIRCUIT OF ARABIDOPSIS CIRCADIAN CLOCK GENE NETWORK

Smirnova O.G.*, Stepanenko I.L.
Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
e-mail: planta@bionet.nsc.ru
**Corresponding author*

Key words: *transcriptional feedback, network, circadian rhythm, clock, signaling*

Motivation and Aim: The endogenous circadian clock regulates many physiological processes related to plant survival and adaptability being a major regulator of daily metabolic processes, hormonal and stress response pathways. In this work, we set the task to analyze the mechanisms of the circadian clock gene network function in *Arabidopsis*, identify key regulatory circuits that determine the circadian clock, and calculate the dynamics of the gene network.

Methods and Algorithms: Gene network was reconstructed on the bases of published experimental data. The package Pajek was used to visualize the relationships between elements of the gene network (http://pajek.imfm.si/doku.php?id=pajek). Synchronous Boolean models have been used to calculate the dynamics of the gene network.

Results: The circadian clock in Arabidopsis is composed of several interconnected feedback loops, coupled morning and evening oscillators. The morning genes CCA1 and LHY directly repress the expression of the evening gene TOC1 during the day, and the accumulation of TOC1 in the evening directly represses the expression of morning genes to form a negative feedback loop in a 24-h period. RVE8, a homolog of CCA1 and LHY, promotes expression of TOC1 and PRR5 towards the end of the subjective day and forms a negative feedback loop with PRR5. Thus, three homologues RVE8 and CCA1 and LHY function close to the circadian oscillator but act via distinct molecular mechanisms. In addition, CCA1 and LHY form a morning negative feedback loop with PRR9, PRR7 and repress CHE expression. BOA binds to the promoter of CCA1 through newly identified promoter binding sites and activates the transcription of CCA1. CCA1 binds to the evening element of the BOA promoter and negatively regulates its expression.

The evening regulatory loop is represented by TOC1, GI, CHE and evening complex EC, composed of ELF3, ELF4, and LUX. Transcription factors TOC1 and CHE repress CCA1 and LHY, expressed in the morning. The binding of the evening complex to the promoters of PRR9 and LUX suppresses their expression.

The ubiquitination pathway has an additional role in circadian rhythm regulation. ZTL is a substrate-binding subunit of an E3 ubiquitin ligase complex. Interaction with GI stabilizes ZTL and is required for the circadian rhythm of ZTL protein. ZTL establishes periodicity of the clock by regulating the proteasome-dependent degradation of TOC1 and PRR5. CCA1 and LHY bind to the promoters of many other clock and clock-regulated genes. CCA1 and LHY recruit the COP10-DET1-DDB1 (CDD) complex to the evening-expressed TOC1 and GI promoters. DET1 possesses transcriptional co-repressor activity that is necessary for CCA1- and LHY-mediated inhibition of TOC1 and GI transcription. LWD1 positively regulates PRR9, PRR5, and TOC1. Within the positive feedback loop formed by LWD1 and PRR9, LWD1 could activate PRR9 by binding with its promoter, whereas the activation of LWD1 by PRR9 is indirect. The network dynamics have been simulated. This model can be useful to test the coherence of experimental data.

Author index

C

D

E

F

G

H

I

J

K

L

M

N

O

P

R

S

T

U

V

W

X

Y

Z

Sk Сколково

ЧТО ТАКОЕ СКОЛКОВО?

Фонд развития Центра разработки и коммерциализации новых технологий занимается созданием уникального для России центра «Сколково». Цель проекта – формирование благоприятных условий для инновационного процесса: ученые, конструкторы, инженеры и бизнесмены совместно с участниками образовательных проектов будут работать над созданием конкурентоспособных наукоемких разработок мирового уровня в пяти приоритетных направлениях.

КЛАСТЕРЫ

В рамках кластера информационных технологий развиваются стратегические направления информационных технологий – от поисковых систем до облачных вычислений. Он объединяет свыше 100 компаний

Компании кластера космических технологий и телекоммуникаций занимаются космическими проектами и развитием телекоммуникационных технологий. Затрагивается множество сфер деятельности – от космического туризма до систем спутниковой навигации. В кластер входит больше 10 компаний.

В рамках кластера биомедицинских технологий поддерживаются и развиваются инновации в области биомедицинских технологий. В кластер входит свыше 90 компаний.

В рамках кластера энергоэффективных технологий поддерживаются инновации и прорывные технологии, нацеленные на сокращение энергопотребления объектами промышленности, ЖКХ и муниципальной инфраструктуры. В кластер входит свыше 80 компаний.

Главной целью компаний кластера ядерных технологий является инновационное развитие ядерных технологий. Компании кластера создают новые продукты для энергетических рынков, разрабатывают новые материалы и проектируют сложные технологические системы. В кластер входит свыше 20 компаний.

ТЕХНОПАРК

Технопарк оказывает инновационным компаниям-участникам проекта всю необходимую поддержку для того, чтобы они успешно развивали свои технологические активы и корпоративную структуру, превращаясь в лидирующих игроков на глобальных рынках. Технопарк реализует эту задачу, привлекая инфраструктуру, ресурсы и другие возможности проекта «Сколково» и его партнеров и превращая их в набор эффективных сервисов, полностью отвечающих потребностям компаний-участников проекта.

УНИВЕРСИТЕТ

Открытый университет Сколково (ОтУС) – часть экосистемы инновационного центра Сколково: источник абитуриентов (магистров и аспирантов) для Университета Сколково, источник стажеров для компаний-партнеров Сколково, источник проектов для бизнес-инкубаторов.

ОтУС – это лекции, мастер-классы, учебные курсы выдающихся мыслителей, ученых и практиков, поддерживаемое и модерируемое сетевое самообразование, стажировки и сезонные школы, в том числе в компаниях-партнерах и ведущих мировых вузах, проектно-образовательные лаборатории.

ГОРОД

Расположенный в 3 км от Москвы на территории 500 га, Сколково станет городом-лабораторией, где внедрят и апробируют разработанные здесь энергоэффективные технологии. В инограде будут проживать около 21 тыс. человек, еще 21 тыс. будет приезжать на работу. Автором генерального плана Сколково является французская компания AREP при участии инженерной компании SETEC и известного ландшафтного архитектора Мишеля Девиня, которая была отобрана из нескольких десятков претендентов.

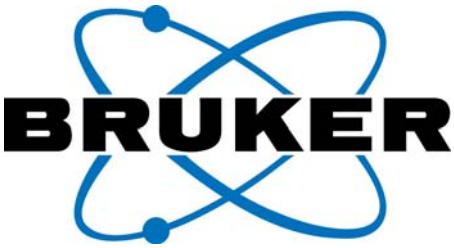
life technologies™

Life Technologies объединяет Applied Biosystems, Invitrogen, Ambion, Gibco, Molecular Probes, Ion Torrent – мировых лидеров в производстве самого современного оборудования и реактивов для широкого спектра научных исследований, молекулярной медицины и ДНК-идентификации личности.

Это глобальная компания, которая предлагает приборы, реактивы, расходные материалы и сервисное обслуживание исследователям по всему миру. Клиентами Life Technologies являются исследователи в области молекулярной и клеточной биологии, репродуктивной и персонализированной медицины, молекулярной диагностики, криминалистики XXI века, а также ученые, работающие в сфере сельского хозяйства и окружающей среды. Продукция Life Technologies используется для поиска и разработки новых лекарств, в токсикологии и криминалистике, для диагностики заболеваний, в клинической клеточной терапии и производстве биопрепаратов.

Компания Life Technologies является сторонником глобальной социальной ответственности. Фонд Life Technologies выделяет миллионы долларов, чтобы сделать мир науки более привлекательным и значимым в глазах общества.

Life Technologies
Российская Федерация, 117485
Москва, ул. Обручева 30/1·
Тел.: +74956516797
Факс: +74956516799
www.lifetechnologies.com



BRUKER is a leading provider of high-performance scientific instruments and solutions for molecular and materials research, as well as for industrial and applied analysis. World's most comprehensive range of scientific instrumentation available under one brand - a brand synonymous with excellence, innovation and quality for 50 years!

Bruker Ltd.
Pyatnitskaya st. 50/2 bld.1
119017, Moscow, Russia
Phone +7 495 5179284, +7 495 5179285
ms@bruker.ru
630090, Novosibirsk, Russia
Prospekt Akademika Lavrentieva 6/1, office 21
Phone: +7 383 333 22 41, +7 383 319 0789
www.bruker.com

Backgrounder

HP in brief

HP

HP creates new possibilities for technology to have a meaningful impact on people, businesses, governments and society. The world's largest technology company, HP brings together a portfolio that spans printing, personal computing, software, services and IT infrastructure at the convergence of the cloud and connectivity, creating seamless, secure, context-aware experiences for a connected world. HP operates in 170 countries, the number of employees being over 324 000 people.

HP in Russia

HP operates on the Russian market for over 40 years. During these years:

- Company's staff increased to more than 1000 employees
- HP's regional network in Russia counts 11 offices throughout the country
- HP constantly expands its investments: opens HP Labs in St. Petersburg, begins software development and starts PC production together with Foxconn
- HP implements a multi-sided Corporate Social Responsibility program in Russia, focusing national priorities such as education and support for culture and scientific research

HP AT A GLANCE

Fortune 11
U.S.

Fortune 28
Global

324,600
employees

$126
BILLION USD
in revenue for FY10

Operates in approximately

170
countries worldwide

© 2011 Hewlett-Packard Development Company, L.P. The information contained herein is subject to change without notice. The only warranties for HP products and services are set forth in the express warranty statements accompanying such products and services. Nothing herein should be construed as constituting an additional warranty. HP shall not be liable for technical or editorial errors or omissions contained herein.

Production association «AMS-MZMO» (OOO «Miasskij Zavod Medicinskogo Oborudovaniya» - ZAO «Asepticheskie Medicinskie Sistemy») is the leading enterprise among Russian organizations in designing and producing of cleanrooms for medical care institutions and differ industries.

«AMS-MZMO» has created and produces as serial equipment the following base elements and components for construction of cleanrooms:

- building envelop:
  - metal framework;
  - wall panels (with or without glass);
  - doors (with or without glass) and windows;
  - hygienic suspended panel and cassette type ceilings;
  - hygienic raster luminescent lamps;
  - material locks and locks for personnel;
  - pass-through windows;
  - antistatic floors;
- systems of air preparing and transfer:
  - systems of positive-pressure ventilation and conditioning;
  - systems of air expenses and pressure control in cleanrooms;
  - air dispensers and air ingress units;
- laminar boxes, air cleaning units, clean areas;
- bandageless burns and wounds treatment units;
- medical gases and power supply consoles;
- equipment control systems;
- modules of biological safety of BSL-4 class.

The quality management system of the enterprises is certificated with reference to design, creation, production, delivery, carrying out erecting and start-up works and certification, repairing and service: medical equipment; complexes of clean and extra clean rooms for electronic, food, medical, microbiological and pharmaceutical industry and medical institutions on compliance to GOST R ISO 9001-2008 (ISO 9001:2000) (ГОСТ Р ИСО 9001-2008 (ИСО 9001:2000)) requirements.

Association of the «AMS-MZMO» enterprises has designed a complex of cleanrooms of a vivarium for small animals (with increased requirements to the contents) for Institute of cytology and genetics of the Siberian Branch of the Russian Academy of Science (Novosibirsk) and as the General contractor carried out its construction during the period from 2005 to 2009. Total area of a vivarium more than 5 800 sq.m, the area of a complex of the cleanrooms which are meeting the requirements of GLP, - 1 000 sq.m.

OOO «MIASSKIJ ZAVOD MEDICINSKOGO OBORUDOVANIYA»

Russia, 456313, Chelyabinsk region, Miass, Turgoyakskoe shosse, 2/16
Tel. +7 (3513) 29-89-01
Fax +7 (3513) 24-25-46
E-mail: laminar@laminar.ru
Internet: http://www.laminar.ru

ОПТЭК
Объединяя решения

1. *Название предприятия, год основания*: Филиал ООО «ОПТЭК» в г. Новосибирске, основан в 2009 году.

2. *Адрес, телефон, в том числе филиалы*: 630090, г.Новосибирск, ул. Инженерная, 28, тел. (383) 363-76-74, 363-76-75

3. *E-mail, сайт*:office-nsk@optecgroup.com, www.optecgroup.com

4. *Лицензия* №99-08-000439 от 10 февраля 2009 года.

5. *Контактные лица для оформления заказов и др.*:
   Морозова Елена Николаевна - директор филиала,
   Деменко Светлана Юрьевна – руководитель направлений офтальмологии и хирургических систем,
   Лапин Андрей Евгеньевич – направление микроскопия и гистология,
   Ломакина Татьяна Юрьевна – расходные материалы.

6. *Специализация предприятия*: торговля медицинским оборудованием.

7. *Ассортимент (наименование, количество позиций, торговые марки, ударные позиции)*: микроскопы, оборудование для моргов, планетарии, офтальмологическое оборудование.

8. *География продаж*: Россия и страны СНГ.

9. *Социальные программы:* спонсорская помощь, гранты на научную деятельность.

ХИМЭКСПЕРТ

Агентство «Химэксперт»
info@khimexpert.ru www.khimexpert.ru
(499) 973-92-80, 972-06-90
127006, г. Москва, ул. Краснопролетарская, д. 7

Официальный дилер AB Sciex и Life Technologies, основной поставщик реактивов и расходных материалов Applied Biosystems и Ambion.

Предлагаем оборудование для анализа ДНК и РНК, фундаментальных протеомных исследований, фармацевтики и биотехнологии, стандартного прикладного тестирования, включая идентификацию личности и установление родства в криминалистике и судебно-медицинской экспертизе, тестирование пищевых продуктов.

- **АМПЛИФИКАТОРЫ**. Applied Biosystems разрабатывает и производит термоциклеры с 1987 года. На нынешний день широкий модельный ряд включает как несложные и недорогие модели, так и новейшую систему, обладающую шестью независимыми температурными зонами, дающими возможность быстро оптимизировать ПЦР и проводить различные ПЦР в одном блоке. Все приборы надежны в использовании, точны и просты в управлении. Вы можете выбрать модель, наилучшим образом отвечающую потребностям вашей лаборатории.
- **СИСТЕМЫ ПЦР В РЕАЛЬНОМ ВРЕМЕНИ.** Applied Biosystems совершенствует технологии систем Real-time PCR уже более 10 лет. Любые современные задачи: анализ экспрессии генов и микроРНК, типирование SNP, выявление транслокаций, определение вирусной нагрузки и т.п. – могут быть решены с применением оборудования Applied Biosystems. Нынешнее пятое поколение систем (ViiA7 и QuantStudio) обладает еще более широким спектром возможностей и рядом уникальных характеристик. Система ViiA7 способна детектировать 21 флуоресцентный краситель, а система QuantStudio позволят проводить цифровую ПЦР (Digital PCR) и работать с OpenArray-слайдами.
- **ГЕНЕТИЧЕСКИЕ АНАЛИЗАТОРЫ**. Applied Biosystems является безусловным лидером в производстве генетических анализаторов (секвенаторов) – специализированных систем капиллярного электрофореза с оптической детекцией флуоресцентного сигнала. Все имеющиеся модели оптимизированы для решения полного спектра задач: определение структуры ДНК, различных типов фрагментного анализа (LOH, AFLP, поиск SNP, STR-генотипирование и др.). Компания предлагает приборы различной производительности: от монокапиллярной системы ABI Prism 310 до 96-ти капиллярной AB 3730xl. Все приборы могут применяться при проведении геномной дактилоскопии (идентификация личности, установление родства), микробиологических, научных и медицинских исследований, а также исследований в сельском хозяйстве.
- **ПРИБОРЫ НОВОГО ПОКОЛЕНИЯ.** Applied Biosystems производит приборы, позволяющие производить полногеномное секвенирование ДНК – SOLiD 5500 и 5500xl с производительностью 90 и 180 млрд. нуклеотидов за запуск соответственно. Это высокопроизводительные приборы, принцип действия которых основан на секвенировании методом последовательного лигирования флуоресцентных олигонуклеотидных проб с последующей регистрацией флуоресцентного сигнала 4-х различных цветов.
  В конце 2012 года появится прибор SOLID 5500W, для которого предусмотрена линейная изотермическая амплификация на поверхности чипа, позволяющая существенное упростить процесс пробоподготовки.
  Также к приборам нового поколения относятся системы Personal Genome Machine™ (PGM) или Ion Torrent и Ion Proton™ Sequencer – приборы для высокопроизводительного секвенирования ДНК без флуоресцентной детекции. Принцип работы: последовательное удлинение олигонуклеотидной затравки ДНК-полимеразой с одновременной регистрацией локального изменения pH в результате встраивания нуклеотидов. Расходный материал: полупроводниковые сенсорные микрочипы с разной производительностью за запуск. С помощью системы PGM можно секвенировать небольшие геномы или отдельные гены с высоким покрытием, система Ion Proton позволит в 100 раз повысить производительность и секвенировать геномы сложных организмов. Несмотря на меньшую, по сравнению с системами SOLiD, производительность, приборы имеют радикальное преимущество перед всеми полно-геномными системами за счет недорогих реактивов.
- **СИСТЕМЫ ВЭЖХ/МС/МС** производства AB Sciex. Спектр предлагаемых систем включает классические тандемные квадрупольные масс-спектрометры, гибридные системы с линейной ионной ловушкой, времяпролетные (TOF и TOF/TOF) «протеомные анализаторы», применяемые для широкого спектра задач – анализ объектов окружающей среды, продуктов питания, фармакологических исследований, криминалистической экспертизы, исследований пептидов и белков, поиска биомаркеров и т.д.

AB SCIEX life technologies™ Ambion® THE RNA COMPANY®

НОЯБРЬ 2011

Средства для разработчиков высокопроизводительных приложений

КОММЕРЧЕСКАЯ ИНФОРМАЦИЯ

ПРОГРАММНЫЕ СРЕДСТВА РАЗРАБОТКИ ДЛЯ ВЫСОКОПРОИЗВОДИТЕЛЬНЫХ КЛАСТЕРНЫХ ПРИЛОЖЕНИЙ

Intel® Parallel Studio XE и Intel® Cluster Studio XE предназначены для разработчиков, использующих языки программирования C++ или Fortran в системах под управлением Windows* и Linux*. Эти средства обеспечивают своевременную разработку кода с высочайшим уровнем производительности и минимальным количеством дефектов на пути кластер-рабочий стол-устройство и дают разработчику возможность рационализации и оптимизации опыта работы конечного пользователя, а также максимального использования возможностей последних многоядерных процессоров Intel®.

Intel® Parallel Studio XE

› Повышение производительности

› Передовые компиляторы C/C++ и Fortran, библиотеки и средства анализа помогают добиться великолепной производительности приложений.

› Intel Parallel Studio XE представляет собой пакет, включающий в себя три продукта:

1. Intel® Composer XE

- Увеличение производительности посредством оптимизации памяти, автоматического параллелизма и векторизации
- Компилятор C/C++ теперь до 47% быстрее по сравнению со своим ближайшим конкурентом
- Компилятор Fortran теперь до 24% быстрее по сравнению со своим ближайшим конкурентом
- Использование Intel® Threading Building Blocks (Intel® TBB) и Intel® Cilk™ Plus дает более широкие возможности реализации параллелизма
- Оптимизации для нового процессора Sandy Bridge

2. Intel® Inspector XE

- Выявляет труднонаходимые ошибки памяти и многопоточности до их возникновения

3. Intel® VTune™ Amplifier XE

- Более интуитивно понятный интерфейс, быстрый график статистических вызовов и просмотр временной линии
- Профилировщик производительности находит узкие места последовательного и параллельного кода, ограничивающие производительность

Intel® C++ Studio XE и Intel® Fortran Studio XE предлагают набор средств, аналогичный Intel Parallel Studio, но только с одним языком.

Набор	Компилятор C/C++	Компилятор Fortran
Intel® Parallel Studio XE	☑	☑
Intel® C++ Studio XE	☑	
Intel® Fortran Studio XE		☑
Intel® Visual Fortran Studio XE		☑
Intel® Cluster Studio XE	☑	☑

Intel® Cluster Studio XE

› Повышение производительности распределенных приложений

› Масштабирование приложений с перспективой на будущее и их более быстрое масштабирование для обеспечения развивающихся вычислительных процессов высокопроизводительных кластерных систем. Масштабирование приложений на мультиядерных процессорах сегодняшнего дня и многоядерных процессорах будущего.

› Intel Cluster Studio XE – первый набор инструментов для разработки высокопроизводительных кластерных приложений, который сочетает в себе передовые средства анализа для использования с MPI и ведущие библиотеки MPI, компиляторы, библиотеки и модели программирования компании Intel.

Масштабируемая производительность — Возможность работы на большем количестве узлов

- Время задержки (латентность) MPI – скорость работы библиотеки Intel® MPI Library до 6,5 раз превышает скорость работы альтернативных библиотек MPI
- Производительность компилятора – ведущие в отрасли компиляторы Intel на языках C/C++ и Fortran

Масштабирование с перспективой на будущее — Поддержка многоядерных процессоров

- Библиотека Intel® MPI Library масштабируется на более чем 90 тысяч процессов
- Модели параллельного программирования – поддерживаемые в промышленном масштабе версии Intel® Threading Building Blocks и Intel® Cilk™ Plus с открытым кодом, MPI, OpenMP, Coarray Fortran
- Цель – сохранение инвестиций в программирование на мультиядерных процессорах для будущего использования на многоядерных машинах

Эффективность масштабирования — Настройка и отладка на большем количестве узлов

- Контроль корректности памяти и многопоточности – теперь Intel® Inspector XE имеет MPI на многих узлах
- Быстрое профилирование производительности на уровне узла – Intel® VTune™ Amplifier XE находит "горячие" точки быстрее при работе на тысячах узлов

ТЕРМИНОЛОГИЯ

Кластер

Компьютерный кластер - это группа, связанных между собой компьютеров, которые так тесно работают друг с другом, что фактически составляют единый вычислительный ресурс. Компоненты кластера обычно именуются "узлами", и они соединены друг с другом.

Компилятор

Компилятор - это программа, которая преобразует текст, написанный на языке программирования, (исходный код) в другую программу на машинном языке (целевой язык). На выходе мы обычно получаем исполняемую программу.

Библиотека Intel® Integrated Performance Primitives (Intel® IPP)

Intel IPP представляет собой набор процедур для мультимедийных и информационно-ориентированных приложений (например, обработка изображений, звука, сигналов, сжатие, криптографические функции).

Библиотека Intel® Math Kernel Library (Intel® MKL)

Intel MKL представляет собой набор математических/числовых функций и процедур (например, матричные операции, быстрые преобразования Фурье (FFT) и т.д.). Библиотека совместима с компиляторами C++ и Fortran, а также другими компиляторами (например, компиляторы Microsoft, GNU Compiler Collection).

ЦЕЛЕВЫЕ ЗАКАЗЧИКИ

Intel Parallel Studio XE предназначен для разработчиков программного обеспечения, работающим в трех основных сегментах:

1. Высокопроизводительные вычислительные системы (HPC)

- Научно-исследовательские центры и университеты
- Прогноз погоды/климатология, оборонная промышленность
- Нефтепоисковые работы/геофизика (сейсмическое моделирование, пластовое моделирование)
- Производство (вычислительная гидромеханика/анализ методом конечных элементов/моделирование автомобилей, самолетов, аэрокосмических аппаратов, устройств и т.д.)
- Финансы (инвестиционные банки/торговля на бирже/управление рисками)

2. Программисты на языке C++ независимых продавцов ПО (Microsoft Visual Studio*/Linux*), для решения следующих вопросов:

- Обработка изображений
- Обработка сигналов
- Компьютерные игры
- Предприятие

3. Промышленные разработчики и производители встроенных систем

- Медицинские устройства
- Производители оборудования

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
postmaster@bio-rad.ru
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)*
- **Bio-Plex 3D™** *(технология «Жидких биочипов», анализ до 500 сиквенс-специфических событий)*

Протеомные технологии *(структурная и функциональная протеомика)*

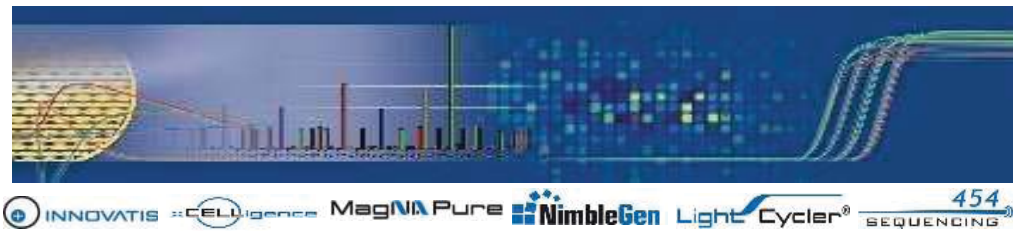
- **BioLogic DuoFlow™** *(модульная гибкая система для биохроматографии)*
- Широкий спектр колонок и носителей
- **Profinia™** *(автоматизированная хроматографическая система очистки рекомбинантных белков)*
- **Pofinity eXact™** *(уникальная бесприборная технология аффинной очистки рекомбинантных белков, свободных от тэг-носителей)*
- **ProteoMiner™** *(уникальная система процессинга белковых комплексов)*
- Системы аналитического и препаративного электрофореза (1-D, 2-D)
- **Rotofor™** (аналитический и препаративный изоэлектрофокус)
- Оборудование для анализа и процессинга 2-D протеомных карт
- **Bio-Plex 200 и Bio-Plex 3D™** *(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*)
- **TC10™** *(слайд-опосредованный счетчик клеток)*

ООО «Рош Диагностика Рус»

Компания ООО «Рош Диагностика Рус» представляет в России швейцарский холдинг F. Hoffmann-La Roche и его подразделение **Roche Applied Science**. Мы предлагаем научно-исследовательским лабораториям инновационное высокотехнологичное оборудование для геномного секвенирования, клеточного анализа, ПЦР в реальном времени и экстракции нуклеиновых кислот в сочетании с наборами высококачественных реагентов. Пакет услуг, предоставляемый компанией, включает инсталляцию приборов, профессиональное сервисное обслуживание, обучение, информационную и методическую поддержку.

www.roche-applied-science.com
www.roche.ru

107031, г. Москва,
Трубная пл., д. 2
Тел. +7 (495) 229-69-99
+7 (495) 229-29-99



Научное издание

Тезисы VIII Международной конференции «Биоинформатика регуляции и структуры генома\ Системная биология»

на английском языке

Abstracts of the Eighth International Conference on Bioinformatics of Genome Regulation and Structure\ Systems Biology

Abstracts have been printed without editing
as received from the authors

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

Дизайн и компьютерная верстка: А.В. Харкевич

Подписано к печати 8. 06. 2012 г.
Формат бумаги 70 × 108 1/16. Печ. л. 32. Уч.-изд. л. 36,5
Тираж 200. Заказ 245

Отпечатано в типографии ФГУП «Издательство СО РАН»
630090, Новосибирск, Морской пр., 2