RUSSIAN ACADEMY OF SCIENCES
SIBERIAN BRANCH

INSTITUTE OF CYTOLOGY AND GENETICS

THE SIXTH INTERNATIONAL CONFERENCE ON BIOINFORMATICS OF GENOME REGULATION AND STRUCTURE

Abstracts

BGRS'2008
Novosibirsk, Russia
June 22—28, 2008

Novosibirsk
2008

INTERNATIONAL PROGRAM COMMITTEE *

Nikolay Kolchanov Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia (Chairman of the Conference)
Ralf Hofestadt University of Bielefeld, Germany (Co-Chairman of the Conference)
Dagmara Furman Institute of Cytology and Genetics SB RAS, Novosibirsk, (Conference Scientific Secretary)
Dmitry Afonnikov Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
Shandar Ahmad National Institute of Biomedical Innovation, Japan
Philip Bourne University of California San Diego, San-Diego, USA
Samir Brahmachari Institute of Genomics and Integrative Biology, Delhi, India
Ming Chen Department of Bioinformatics Zhejiang University, Hangzhou, China
A. Fazel Famili University of Ottawa, IIT/ITI - National Research Council Canada, Ottawa, Canada
Mikhail Gelfand Institute for Information Transmission Problems RAS, Russia
Boris M. Glinsky Institute of Computational Mathematics and Mathematical Geophysics SB RAS, Novosibirsk, Russia
Nikolay Goncharov Institute of Cytology and Genetics, Novosibirsk, Russia
Charlie Hodgman Multidisciplinary Centre for Integrative Biology, School of Biosciences, University of Nottingham, UK
Alexis Ivanov Institute of Biomedical Chemistry RAMS, Moscow, Russia
Manfred Kayser Erasmus University Medical Centre Rotterdam, Rotterdam, The Netherlands
Alexey Kochetov Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
Fyodor Kondrashov Division of Biological Sciences, University of California at San Diego, USA
Jun-ichi Kudoh Center for Northeast Asian Studies, Tohoku University, Japan
Vladimir Kuznetsov Division of Genome and Gene Expression Data Analysis, Bioinformatics Institute, Singapore
Mikhail Lavrentiev Sobolev Institute of Mathematics, Institute of Automation and Electrometry, Novosibirsk, Russia
Sergei Lukaschuk Hull Institute for Mathematical Science and Applications, The University of Hull, UK
Luciano Milanesi National Research Council - Institute of Biomedical Technology, Italy
Kenji Mizuguchi Bioinformatics and Computational Biology Group at the NiBio, The National Institute of Biomedical Innovation, Japan
Mikhail Moshkin Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
Vladimir Poroikov Laboratory for Structure-Function Based Drug Design, Institute of Biomedical Chemistry of RAMS, Moscow, Russia
Jagath Rajapakse School of Computer Engineering, Nanyang Technological University, Singapore
Igor Rogozin National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, USA
Nikolay Rubtsov Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
Maria Samsonova St.Petersburg State Polytechnic University, St.Petersburg, Russia
Akinori Sarai Kyushu Institute of Technology (KIT), Iizuka, Japan
Konstantin Skryabin "Bioingineering" Center, RAS, Russia
Natalia Sournina Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
Evgenii Vereshagin Siberian Center of pharmacology and biotechnology, Novosibirsk, Russia
Evgenii Vityaev Sobolev Institute of Mathematics, Novosibirsk, Russia
Valentin Vlassov Institute of Chemical Biology and Fundamental Medicine of SB RAS, Novosibirsk, Russia
Lev Zhivotovsky Institute of General Genetics RAS, Moscow, Russia

LOCAL ORGANIZING COMMITTEE

Ilya Akberdin Institute of Cytology and Genetics, Novosibirsk, Russia (Chairperson)
Kirill Bezmaternykh Institute of Cytology and Genetics, Novosibirsk, Russia
Ekaterina Denisova Institute of Cytology and Genetics, Novosibirsk, Russia
Luba Glebova Institute of Cytology and Genetics, Novosibirsk, Russia
Nadezda Glebova Institute of Cytology and Genetics, Novosibirsk, Russia
Tatyana Karamisheva 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
Sergey Lavryushev Institute of Cytology and Genetics, Novosibirsk, Russia
Anna Onchukova Institute of Cytology and Genetics, Novosibirsk, Russia
Svetlana Zubova Institute of Cytology and Genetics, Novosibirsk, Russia

ISBN 978-5-91291-005-0

Organizers

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

Siberian Branch of the Russian Academy of Sciences

The Vavilov Society of Geneticists and Breeders

Laboratory of Theoretical Genetics

Novosibirsk State University

Chair of Information Biology

PBSoft Ltd.

Institute of Computational Technologies

Sponsors

Russian Foundation for Basic Research

Carl Zeiss

Bio-Rad

Scientific Professional Equipment

Siberian Branch of the Russian Academy of Sciences

Federal Agency for Science and Innovation

Hewlett Packard

HTLabAG

Sections of the Conference

1. Genomics & Transcriptomics
2. Proteomics
3. Computational Systems Biology
4. Evolutionary and Population Computational Biology
5. Intelligence Data Analysis and Pattern Recognition in Bioinformatics
6. Bioinformatics and New Pharmacology
7. Computer Analysis and Image Recognition in Systems Biology
8. High-Performance Computing in Bioinformatics
9. Nanobioengineering
10. Microsymposium “Genetic Collections and Biodiversity of Cultivated Plants: Producing, Preservation and Cryoconservation”
11. First Micro-Satellite Symposium “Genetic Models in Postgenomic Biology”

Abstracts have been printed without editing as received from the authors

Contents

COMPARISON ANALYSING OF MUTANT GENE *cbn1* WITH MUTANT GENE *cao* AND MOLECULAR MAPPING OF *CAO* GENE IN CHLAMYDOMONAS REINHARDTII

Abdulla H.[1], Mijit Gh.[1 *], Xu qin [1], Rahman E.[1], Chunaev A.S.[2]
[1]Departmen of Cytology and Genetics of Xinjiang University,Urumqi, China
[2]Department of Genetcs and Breeding, St.Petersburg University, St.Petersburg, Russia
E-mail:ghopurm@xju.edu.cn

Motivation and Aim: The unicellular, green alga *Chlamydomonas reinhardtii* has many characteristics that make it an ideal organism for elucidating the function, biosynthesis, and regulation of the photosynthetic pigments and apparatus. For the moment, it is still uncertain whether the Chlamydomonas mutant gene *cbn1* and the mutant gene *cao* without the ability of chlorophyll b synthesis are allelic genes, and the issue of their molecular mapping is also to be settled. This article is dedicated to reporting research work on the above issues.

Methods and Algorithms: (1) Transformation recovery test: plasmid Psp109-E8 with *CAO* was transformed into *Chlamydomonas cbnl-48 mt+* mutant strain by electroporation, recovery expression of chlorophyll b was detected, which verified primarily that *cbn1* and *cao* without the ability of chlorophyll b systhesis both possess allelic characters. (2)Genetic analysis test: mutant strains, 1641-1b (*cbn1-43 mt+*) and CC-1354 (*cbn1-48 mt+*), were taken to mate with segregants of mutant strain CBS5-c1 (*cao5 mt-*) and CBS5-c5(*cao5 mt-*), 1842 meiotoc segregants were obtained by random cell mating but no wild type segregants was found this proved the closely linkage between mutant gene *cbn1* and *cao*. And among the 50 segregants obtained by cell mating tetrad analysis no wild type segregants was detected, which further verified the allelic characters of mutant gene (3) DNA blot test: design two probes according to *CAO* sequence, and carry out DNA dot blot with 21 BAC clones between the two molecular markers, GBP1 and RB47, on *chalamydomonas* linkage group I. The experimental result showed that two probes possess the homology region with 21th BAC clones (BAC number: 33e2). This result prooved that *CAO* exactly locate at cbn1 loci.

Conclusion: Data obtained in experiments for progeny random analysis prove the closely linkage between the mutant genes, *cbn1* and *cao*. Experimental results of genetic tranformation prove that mutant genes, *cbn1* and *cao* possess allelic character.Experimental results of DNA dot blot analysis proved that the genetic loci of *CAO* is near cbn1 and in the fragment region of 33e2 BAC clone which is between the two molecular markers, GBP1 and RB47.

Acknowledgments: We thank professor Ayumi Tanaka (from Hokkaidu University, Japan) for providing some mutants and plasmid to us for this research.We also feel grateful to Dr. E.Harris (from American Duke University) for her gracious support and help.

(This work was supported by the China National Science Foundation (grant no.30060035, 30470914)

References:

1. О.Н.Мирная, Ю.Г.Фомина –Ещенко, А.С.Чунаев.（1990). Локализация мутации *cbn1* в первой группе сцепления ядерных генов *Chlamydomonas reinhardtii*. *Генетика*,26(5):958-960
2. A.Tanaka et al.(1998). Chlorophyll *a* oxygenase (*CAO*) is involved in chlorophyll *b* formation from chlorophyll *a*, *Plant Biology*, Vol.95,12719–12723

SUPERVISED LEARNING APPROACH TO IMPROVE ERROR PROBABILITIES FOR SHORT READ DNA SEQUENCING

Abnizova I.*, Whiteford N., Skelly T., Brown C.
Wellcome Trust Sanger Institute, Hinxton, UK, e-mail: ia1@sanger.ac.uk
** Corresponding author*

Aim and Motivation: Around 10 years ago the *Phred* [1] base calling and error calibration algorithm was introduced. It is threshold dependent binning greedy (nearly optimal) algorithm, allowing estimation of an error probability for every base call given a number of its parameters. As far we know, since then there were no significant attempts, except of [2], to improve the *Phred* error probability calibration.

The new short read sequencing technique introduced new technological and computational challenges. It requires some re-considering of well-known error estimation algorithm taking into account different sequencing platforms. Overall, *Phred* is a fair predictor. However, at high quality bases (*Phred* Q> 30) it tends to under-predict the error rate.

Methods and Algorithms: Because the processes causing errors in a base calling are still not absolutely clear and well defined, a reasonable approach to handle an error prediction will be learning from the data. To calibrate the data, we suggested to use a supervised learning techniques because of their optimization flexibility and high predictive power, as well as ability to deal with a multi-dimensional space and non-linear dependencies between its elements. We suggest to use a special bootstrap-type sampling techniques to deal with typical machine learning problem, namely, with imbalanced training set.

Results: Due to pilot results, we managed to produce 60% of sequences with quality as high as *Phred* Q30.

Availability: The algorithm is implemented into automatic Solexa/Illumina pipeline.

References:

1. Ewing, P. and Green, P. (1998) Base-calling of automated sequencer tracer using Phred. II. Error probabilities. *Genome Research* 8:186:194
2. Brockman, W. et al. (2008) Quality scores and SNP detection in sequencing-by-synthesis detection. *Genome Research* doi 10.1101/gr.070227.107

MATHEMATICAL MODEL OF AUXIN METABOLISM IN SHOOTS OF *ARABIDOPSIS THALIANA L*

Akberdin I.R.[1]*, Omelyanchuk N.A.[1], Fadeev S.I.[2, 3], Efimov V.M.[1, 3], Gainova I.A.[2], Likhoshvai V.A.[1, 3]

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

Motivation and Aim: Auxin participates in regulation of cell differentiation in development of embryo, leaves, vascular tissue, fruit, primary and lateral root and in controlling apical dominance and tropisms. Indole-3-acetic acid (IAA) is physiologically active in the form of the free acid, but can also be found in conjugated forms in plant tissues. IAA can be degraded and redundant pathways lead to its synthesis. The regulation of the IAA metabolism (synthesis, conjugation and degradations) is enough complex and may explain in some aspects how this simple substance is able to influence such diverse processes. Mathematical modeling of IAA metabolic gene network can help reveal the main factors governing this complex process;

Methods and Algorithms: To reach this aim, we first reconstructed a gene network of auxin biosynthesis, conjugation degradation by annotating experimental data from 105 published papers into GeneNet computer system [1]. This gene network after reduction was input into converter [2] to generate the mathematical model of auxin metabolism;

Results: The gene network for the auxin metabolism contains 62 genes and 44 proteins. Different molecular genetic processes within IAA metabolism network take place in 7 compartments. Within this network one regulatory contour with positive feedback and one regulatory contour with negative feedback were identified. For model generation the main regulatory factors in this gene network were selected according to consequences of their disruption for the network maintenance and plant organism functioning. By this way the whole complex network was reduced to core gene network. Based on the mathematical model of the core gene network, we followed the dynamics of the developed gene network and analyzed the key regulatory patterns. We reproduce the experimental data on IAA content in arabidopsis seeds [3].

Conclusion: We have reconstructed the gene network and develop the mathematical model of auxin metabolism in arabidopsis shoots. The model allows to reproduce some phenomenological and molecular-genetic aspects of the auxin role in the plant development.

References:

1. Ananko E.A et al. (2005) GeneNet in 2005. *Nucleic Acids Res.,* **33**: 425-427.
2. Kazancev F.V et al. (2007) Automatical generation of molecular genetic system mathematical models on basis of gene networks structure, *this issue.*
3. Park S. et al. (2001) Partial characterization of major amide-linked conjugates of IAA in Arabidopsis seed (Abstract #321). Final Program July 2001, American Society of Plant Biologists/ Canadian Society of Plant Physiologists meeting, Providence, Rhode Island: 81-82.

OPENMP+MPI PARALLEL IMPLEMENTATION OF THE "MOLKERN" MOLECULAR MODELLING SOFTWARE PACKAGE

***Alemasov N.A.*[1], *Fomin E.S.*[2*]**
[1] Novosibirsk State University, Novosibirsk, Russia
[2] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
e-mail: fomin@bionet.nsc.ru
* Corresponding author

Motivation and Aim: In recent years, new approaches using molecular mechanics replaced traditional score functions in solving molecular docking and virtual screening tasks. These approaches, in spite of being more computationally time-consuming, have considerable advantages: high precision of results and applicability to all classes of chemical compounds. MOLKERN, a software package developed in IC&G SB RAS [1], implements the basic elements of molecular modeling, docking and virtual screening tasks without using any score functions. The package includes algorithms with the computational scaling no greater than O(N log N).

At present, MOLKERN is being developed to be run on supercomputing clusters. It uses the MPI library for data transfer between processors. The MPI is used for parallel performing of tasks of the same type, for instance, virtual screening of chemical compounds libraries. The program module – parallel processes dispatcher – has been integrated in the MOLKERN for efficient solution of such tasks.

For parallel running of tasks that cannot be divided into sets of same and independent subtasks, the OpenMP parallelizing technology is used. This technology is applied to distributing computations among cores of same processors. To solve the common data access problem without performance loss, "mirror" arrays are created for all data used in the write mode. On completion of every thread, all output arrays are merged. This approach doesn't use any critical sections, which allows obtaining significant calculation speedup.

Results: A computer with a dual-core AMD Athlon X2 5000+ EE processor, 1 GB RAM, MS Windows XP operating system, and MS Visual Studio 2005 compiler was used for test calculations. The new OpenMP + MPI version of MOLKERN demonstrates the following speedup characteristics:

Minimal speedup: 18%, 1AIE complex, 522 atoms, 7 threads.
Maximal speedup: 48.4%, 1GC1 complex, 14104 atoms, 2 threads.
Average speedup: test set of 5 complexes (522-14104 atoms): 40.4%.

References:

1. E.S.Fomin, N.A.Alemasov, A.S.Chirtsov, A.E.Fomin (2007) MOLKERN as new effective engine for drug discovery software, *4th International Symposium Computational Methods in Toxicology and Pharmacology Integrating Internet Resources*: 97.

FUNCTIONAL ANNOTATION OF AMINO ACID SEQUENCES USING THE LOCAL SIMILARITY

Alexandrov K.E.*, Sobolev B.N., Filimonov D.A., Poroikov V.V.
The V.N. Orekhovich Institute of Biomedical Chemistry of Rus. Acad. Med. Sci., Moscow, Russia; e-mail: dzimmu@yandex.ru
* Corresponding author

Motivation and Aim: The functional annotation of amino acid sequences is one of the most important problems of bioinformatics. Different approaches were successfully applied for recognition of some functional classes; nevertheless for many functional groups predication still cannot be obtained with the required accuracy. The aim of our study is to develop a method for recognition of protein function, which might be applied to broad classes of proteins.

Methods and Algorithms: We developed a machine learning method of the protein function recognition based on the original method of sequence description. All sequences from the training set are compared with the sequence under study (query sequence), by shifting them along the query sequence. So the query sequence is represented by the set of local similarity scores for all its positions. These scores are used as the input data for the original classifier based on the Bayesian approach, which estimates the probability of the annotated protein belonging to the considered functional classes. The prediction accuracy was estimated by the leave-one-out cross-validation (LOOCV) procedure. In order to estimate the comparative efficiency of suggested method we executed also functional class prediction with the methods implementing the Hidden Markov Models (HMM) and Support Vector Machine (SVM).

Results: At first, the method was tested versus two sets covering 28 serine protease classes and 56 Gold Standard families, respectively. We showed that our method predicts effectively the functional class of proteins when these classes do not intersect each to others. The level of prediction accuracy was 100% for 45 from 56 Gold Standard families. We also tested the method with cytochrome P450 superfamily partitioned into the intersecting classes of substrate or inducer specificity. In this case the accuracy of prediction for some classes was significantly lower — about 14%. However, the accuracy increased according to increasing number of the class members. The acceptable level of prediction accuracy (exceeded 60%) was reached for intersecting classes with at least 10 members. Also 100% accuracy was reached for some intersecting classes. The HMM-based program showed the accuracy comparable with the accuracy of our method, while the results obtained with SVM were less accurate.

Conclusion: Proposed approach revealed 100% recognition for the majority of non-intersected classes and, thus, it can be used for functional mapping of amino acid sequences.

Reference:

1. Alexandrov K., Sobolev B., Filimonov D., Poroikov V. (2008) Recognition of protein function using the local similarity. J. Bioinform. Comput. Biol. In press.

SYSTEMS BIOLOGY STUDIES OF THE EFFECTS OF LEPTIN REPLACEMENT ON HUMAN GLUCOSE HOMEOSTASIS

Andreev V.P *, Paz-Filho G., Wong M-L., Licinio J.
Center for Pharmacogenomics, Department of Psychiatry and Behavioral Sciences, University of Miami Miller School of Medicine, Miami, FL 33136, USA
e-mail: vandreev@med.miami.edu
* Corresponding author

Motivation and Aim: Obesity is a major public health problem worldwide. Leptin plays a central role in the regulation of food intake and energy expenditure. Studying the metabolism of leptin-deficient patients before and after leptin replacement treatment helps to clarify the mechanism of leptin action.

Methods and Algorithms: The only genetically leptin-deficient adult man identified in the world to date was treated for 24 months with recombinant methionyl human leptin. Blood was collected every 7 min for 24 hours at baseline, one-week, 18-months and 24-months after initiation of the treatment. Concentrations of insulin, C-peptide and plasma glucose were measured. Insulin secretion throughout the day was obtained by de-convolution of C-peptide data. Hepatic extraction was estimated based on our modification of the insulin kinetics model. Insulin sensitivity for each of the four meals was calculated by using the meal tolerance test approach and minimal glucose model.

Results: Hepatic extraction of insulin was the first element of glucose homeostasis to respond to leptin replacement treatment and increased 2-fold after one week of treatment. Insulin secretion and delivery rates decreased more than 2-fold after 18 months of treatment. A several-fold increase in insulin sensitivity was observed after both 18 and 24 months of leptin replacement treatment. Therefore, it might be concluded that leptin replacement acutely increases insulin sensitivity and insulin hepatic extraction without short-term effect on insulin secretion. Insulin sensitivity is further increased 18 months after leptin replacement, with a decrease in insulin secretion and normalization of hepatic extraction. After 24 months, insulin sensitivity further increases and insulin secretion slightly decreases. This suggests a time-dependent effect of leptin on the glucose homeostasis that culminates with increased insulin sensitivity and decreased insulin secretion in the long-term.

Conclusion: Results of the current experiments on quantitative proteomics analysis of our patient's plasma and adipocyte samples "before" versus "after" and "on leptin" versus "off leptin" will be presented. Approaches to create multiscale systems biology models by combining phenomenological models at the organism level with the pathway models at the cellular level will be discussed.

Availability: In-lab developed set of programs, MIGS (Miami Insulin Glucose Software) is written in MATLAB and available from the authors on request.

HIV-1 GP120 V3-LOOP COMPARATIVE STRUCTURE ANALYSIS: SEARCH FOR THE STRUCTURALLY CONSERVED REGIONS

Anishchenko I.V.
Belarusian State University, Minsk, Republic of Belarus
e-mail: anishchenko.ivan@gmail.com

Motivation and aim: The object of the current study is the third hypervariable V3 region of the HIV-1 gp120 protein, which is responsible for many aspects of viral infectivity. It is remarkable for its sequence diversity and, hence, structural diversity, which brings sufficient complications in the study of the V3-loop. At the same time information on the V3-loop conserved regions within its preferred conformations could be a significant tool for anti-AIDS drug design. On account of the V3-loop sequence diversity the conservative sequences of the HIV-1 group M subtypes are taken into consideration. On the basis of the V3-loop structures of definite isolates obtained before 3D modeling for each subtype sequence is performed. The study then is sighted at the search for structurally invariant regions in the models obtained, which may be regarded as drug targets.

Methods and Algorithms: 3D structure modeling is performed in MODELLER, also supporting functions of de novo modeling of loops in protein structures (http://www.salilab.org/modeller/). Several V3-loop structures obtained from the NRM and X-ray studies are taken as templates [1-2]. The consensus sequences of the V3-loop, being targets for modeling, are supported by the HIV Sequence Database at Los Alamos (http://www.hiv.lanl.gov/content/sequence/HIV/mainpage.html). Structural analysis is held in Bio3d package for the R environment (http://mccammon.ucsd.edu/~bgrant/bio3d/). Both comparison in the geometric spaces of Cartesian coordinates and dihedral angles are performed.

Availability: In the result of the study we pay attention to the regions, which preserve their conformational states within the structures under review. The most probable ones are drawn out to be a starting point for further drug design.

References:

1. A.M. Andrianov, V.G. Veresov (2006) Determination of Structurally Conservative Amino Acids of the HIV-1 Protein gp120 V3 Loop as Promising Targets for Drug Design by Protein Engineering Approaches, *Biochemistry (Moskow)*, **71**: 906-914
2. Huang C.-C., M. Tang, M. Y. Zhang, S. Majeed, E. Montabana, R. L. Stanfield, D. S. Dimitrov, B. Korber, J. Sodroski, I. A. Wilson, R. Wyatt, and P. D. Kwong (2005) Structure of a V3-containing HIV-1 gp120 core, *Science*, **310**: 1025–1028

TEpredict: SOFTWARE FOR PREDICTING T-CELL EPITOPES

Antonets D.V.*, Maksyutov A.Z.
SRC VB "Vector", Novosibirsk region, Koltsovo, Russia
e-mail: antonec@vector.nsc.ru
* Corresponding author

Motivation and Aim: T-cell epitopes are important tools for diagnosis and treatment of infectious, autoimmune or cancer diseases as well as for the development of polyepitope vaccines. An accurate prediction of T-cell epitopes is an urgent task of bioinformatics within immunology.

Methods and Algorithms: Models for predicting MHC-peptide binding affinity were constructed by means of Partial Least Squares regression using quantitative MHC-peptide binding data, collected from IEDB (http://www.immuneepitope.org). We used pls package [1] for R (http://www.r-project.org). Performance of generated models was assessed with ROCR package [2]. All programs were written in Python programming language; Python interacted with R through the RPy interface (http://rpy.sourceforge.net/). Predictions with TEpredict could be alternatively done using models implemented in ProPred1 [3] and ProPred [4] web-servers (for predicting MHC I- or MHC II-binders respectively). Predictions of proteasomal and immunoproteasomal processing were implemented using models originated from ProPred1. For predicting peptide-TAP binding we used models taken from the literature sources [5, 6].

Results: Our TEpredict software could be used for predicting MHC I- as well as MHC II-binders. It allows to predict proteasomal processing and peptide-TAP binding. TEpredict can exclude peptides shearing local similarity with human proteins from the set of predicted epitopes. It is able to estimate expected population coverage by selected set of peptides using HLA allele frequencies data taken from (http://www.ncbi.nlm.nih.gov/gv/mhc/). Besides graphical user interface TEpredict have command line interface or it could be used as python package. This makes possible to automate analysis of huge amounts of data. Comparative testing with some other widely used predictive methods showed TEpredict to be highly competitive or to outperform them.

Availability: TEpredict is available on request from the authors and for academic organizations is free of charge.

References:

1. Mevik B.-H., Wehrens R. (2007) The pls Package: Principal Component and Partial Least Squares Regression in R. Journal of Statistical Software, 18: 1-24.
2. Tobias S. et al. (2005) ROCR: visualizing classifier performance in R. Bioinformatics, 21: 3940-3941.
3. Singh H., Raghava G.P. (2003) ProPred1: prediction of promiscuous MHC Class-I binding sites. Bioinformatics, 19: 1009-1014.
4. Singh H., Raghava G.P. (2001) ProPred: prediction of HLA-DR binding sites. Bioinformatics, 17: 1236-1237.
5. Peters B. et al. (2003) Identifying MHC class I epitopes by predicting the TAP transport efficiency of epitope precursors. J Immunol, 171: 1741-1749.
6. Doytchinova I. et al. (2004) Transporter associated with antigen processing preselection of peptides binding to the MHC: a bioinformatic evaluation. J Immunol, 173: 6813-6819.

DEVELOPMENT OF THE COMPUTER PROGRAM FOR DEFINING LEAF HAIRINESS IN WHEAT BASED ON ITS MICROSCOPE IMAGE PROCESSING

***Arsenina S.I.*[1]*, *Afonnikov D.A.*[1, 2], *Pshenichnikova T.A.*[2]**
[1] Novosibirsk State University, Novosibirsk, Russia
[2] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
e-mail: siata@mail.ru
* Corresponding author

Motivation and Aim: Leaf hairiness in wheat plants is of great importance for protection from pests and for adaptation to environmental factors. For example, this trait is characteristic of a number of drought resistant wheat cultivars referred to the steppe ecological group. To identify the genes responsible for the leaf hairiness, mass analysis of a great number of plants belonging to different hybrid populations is needed, accompanying with a laborious manual job. Furthermore, a more accurate description of the morphological properties of the trait for correct determination of phenotypic classes is timely. In the course of the work, we developed the LHDetect program for determining the degree of leaf hairiness and its morphological properties on the basis of its microscope image processing.

Methods and Algorithms: Equipment of the General Usage Center for Microscopic analysis SB RAS was used to obtain microscope images. To determine hairiness, images of the leaf cross fold were obtained at 5-fold magnification. Optimal conditions were chosen for leaf positioning, leaf and background illumination for obtained microscope images.

The algorithm for determination of the hairiness degree was two-step:

1). Definition of the leaf boundary on the image.

2). Counts of the number of trichomes at different distances from the leaf boundary.

Results and Conclusion: The algorithm for defining of the hairiness degree of wheat leaf was implemented in Java. Preliminary tests were carried out to estimate the algorithm performance. For example, the mean absolute error deviation between the automatically detected leaf boundary and the boundary determined “by an eye” is estimated as 18.25 pixels per 1030 pixels of the boundary length.

NESTED GENES AND THE EVOLUTION OF METAZOAN GENOME ORGANIZATIONAL COMPLEXITY

Assis R.[1], Kondrashov A.S.[1], Koonin E.V.[2], Kondrashov F.A.[3]*
[1]Center for Computational Medicine and Biology and the Life Sciences Institute, University of Michigan, Ann Arbor MI, 48109 USA.
[2]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda MD, 20894 USA.
[3]Section on Ecology, Behavior and Evolution, Division of Biological Sciences, University of California at San Diego, 2218 Muir Biology Building, La Jolla CA, 92093 USA.Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
e-mail: kondrashov@ucsd.edu
* Corresponding author

Motivation and Aim: It is conventionally assumed, that the course of evolution proceeds without any predetermined directionality, such that only selection and mutation are the main determinants of the evolutionary trajectory. However, in some instances evolutionary history provides constraints on the further direction and trajectory of the evolutionary process. We study the direction of evolution of genome organizational complexity, focusing on the evolution of nested gene structures, where one, or several, internal genes are located within another, external gene. We have chosen nested genes as the object of study of organizational complexity because they represent an inherently complex structure, in which genes are arranged in a non-linear fashion.

Methods and Algorithms: We used a comparative genomic approach using completely sequenced genomes from the vertebrate, nematode and fruit fly lineages. Firstly, we identified all nested gene structures from *H. sapiens*, *D. melanogaster*, *C. elegans* and *C. briggsae* genomes. Then, we used sequence comparisons with BLAT and BLASTP algorithms to find orthologs of the nested genes in a sister species, and determined the ancestral state of the nested gene structure by analyzing an outgroup species in a parsimony approach.

Results: We identified 128, 792, 429, and 233 annotated nested gene pairs in H. sapiens, D. melanogaster, C. elegans, and C. briggsae genomes, respectively. Of these, we were able to distinguish 56 gains and 0 losses of the nested state, 48 gains and 15 losses, and 24 gains and 2 losses, respectively, in these 3 lineages. The intriguing pattern of preferred gain of nested genes and the associated increase of genome organizational complexity appears to occur at a relatively uniform rate throughout evolution. We estimate that the rate of gain of nested genes is between 0.2 and 0.9 gains per million years.

Conclusion: It appears that the complexity of genome organization has been independently increasing in at least three metazoan lineages. The slow rate of nested gene structure acquisition is consistent with a constant, independent gain of nested genes since the expansion of spliceosomal introns in the common ancestor of eukaryotes ~1 billion years ago. Thus, the intronless state of genes in the common eukaryotic ancestor continues to drive the increase of genome organizational complexity in extant metazoan genomes.

A BAYESIAN APPROACH TO EVOLUTIONARY HISTORY OF THE FAMILY *POXVIRIDAE*

Babkin I.V.*, Shchelkunov S.N.
State Research Center of Virology and Biotechnology Vector, Koltsovo, Novosibirsk oblast, Russia
e-mail: babkin@vector.nsc.ru
* Corresponding author

Motivation and Aim: The comparison of poxvirus genome with the other viruses suggested the existence of a common ancestor for many of them. Nonetheless, these phylogenetic reconstructions have not been compared with the time scale of virus evolutionary history. Taking into account that variola virus (VARV) was imported to South America from West Africa in the XVI century, we succeeded in assessing the time point whereat the *minor alastrim* and West African VARV strains diverged. Based on this value, we conducted an evolutionary analysis of a wide range of orthopoxvirus and poxvirus strains.

Methods and Algorithms: Phylogenetic analysis was performed by the programs Clustal X, BioEdit, Modeltest, Paup and Mega. The Bayesian dating method was performed by the programs Paml and Multidivtime.

Results: Taking advantage of the unique situation with the known dating of the VARV transfer from West Africa to South America and our own data on the close phylogenetic relationships between the modern West African and South American VARV isolates, we have analyzed the extended central conservative region of the orthopoxvirus genome and eight genes of multisubunit RNA polymerase of various genera from the family *Poxviridae* and calculated for the first time the rate of mutation accumulation in these DNA viruses, which amounts to 10^{-6} nucleotide substitutions per site per year. The divergence of poxviruses from an ancestor virus into the modern genera occurred over 200 Tya; the precursor of the genus *Orthopoxvirus* diverged 131±45 Tya. The rest genera of mammalian poxviruses, having DNA with a low G+C content, diverged approximately 110-90 Tya. VARV started its independent evolution 3.4±0.8 Tya. It has been discovered by the example of VARV and monkeypox virus that the geographic conditions that allowed for isolation of the animals in West Africa from the rest territory of this continent led to a divergent evolution of the orthopoxviruses in question and formation of the West African VARV and MPXV subtypes.

Conclusion: The origination of VARV ancestor, dated back to about 3-4 Tya, demonstrates that VARV is a relatively young virus; this explains the absence of records about smallpox outbreaks in ancient historical records (Talmud, Bible, etc.). The first reliable descriptions of this disease date back to the IV century AD. The calculated rates of molecular evolution of poxviruses are approximately by one-two orders of magnitude lower as compared with the viruses with short DNA genome and by three-four orders of magnitude lower than in the viruses with single-stranded RNA genome; however, they are by three–four orders of magnitude higher compared with the evolutionary rate of animal chromosomal genes. The calculated rate of genetic variation reflects objectively an essentially higher rate of alternation of poxvirus generations compared with their natural hosts, wherein they reproduce. The performed molecular dating of orthopoxviruses demonstrates clearly that the evolutionary potential of cowpox virus is underestimated. VARV had evolved from a cowpox virus-like ancestor virus over a relatively short historical time period. The existing natural resources of cowpox virus, whose main host is rodents, are yet insufficiently studied. The results reported demonstrated that we pioneered in determination of the time scale of divergence evolution in the family *Poxviridae* and evaluation of the rate of mutation accumulation by certain members of this family.

ACCELERATED ADAPTIVE EVOLUTION ON A NEWLY FORMED X CHROMOSOME

Bachtrog D.*, Jensen J.D., Zhang Z.
Division of Biological Sciences, University of California, San Diego, 9500 Gilman Drive, MC 0116, La Jolla, CA 92093
e-mail: dbachtrog@ucsd.edu
* Corresponding author

Sex chromosomes originated from ordinary autosomes, and their evolution is characterized by a continuous gene loss on the ancestral Y chromosome. Here, we demonstrate another principle of sex chromosome evolution: bursts of adaptive fixations on a newly formed X chromosome. Taking advantage of the recently formed neo-X chromosome of *Drosophila miranda*, we compare patterns of DNA sequence variation at 150 gene fragments located on the neo-X to 110 gene fragments on the ancestral X chromosome. This contrast allows us to draw inferences of selection on a newly formed X chromosome relative to background levels of adaptation, while controlling for demographic effects. Chromosome-wide diversity on the neo-X is reduced twofold relative to the ancestral X, as expected under recent and recurrent directional selection. Several statistical tests employing various features of the data consistently identify 10-15% of neo-X genes as targets of recent adaptive evolution but only 1-3% of genes on the ancestral X. A likelihood method to estimate parameters of recurrent adaptive evolution reveals that both the rate of adaptation and the fitness effects of adaptive substitutions are roughly an order of magnitude higher for neo-X genes relative to genes on the ancestral X. Thus, we propose that newly formed X chromosomes are not passive players in the evolutionary process of sex chromosome differentiation but respond adaptively to both their sex-biased transmission and to Y-chromosome degeneration, possibly through demasculinization of their gene content and the evolution of dosage compensation.

ANALYSIS OF THE EVOLUTIONARY SPECIFICITIES OF THE YFIA И YHBH E. COLI GENES AND THEIR REGULATORY REGIONS

Baryshev P.S.[1]*, Oschepkov D.Yu.[1], Khlebodarova T.M.[1], Afonnikov D.A.[1, 2]
[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
[2] Novosibirsk State University, Novosibirsk, Russia
e-mail: pbaryshev@bionet.nsc.ru
* Corresponding author

Motivation and Aim: The YfiA and YhbH *E.coli* proteins belongs to the sigma-54 modulation protein family and their sequences are about 40% similar. These proteins have opposite effect on protein activity affecting the formation of the 70S dimmer particle [1]. It is known that the YfiA gene is activated in the stress exposed cell. This can be taken advantage of in development of a polyfunctional genetic sensor of oxidative stress, specific damage of DNA structure, and nonspecific damage of protein and membrane structure [2]. In this study we analyze bacterial whole genome sequences to identify evolutionary specificities of the YfiA and YhbH genes and their regulatory regions.

Methods and Algorithms: Data on 663 bacterial genome sequences were taken from the NCBI site (ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/). The BLAST program was used to search the homologs of the YfiA and YhbH sequences. Multiple sequence alignment was done using Mafft [3], phylogenetic tree was built by the Phyml program. The regulatory regions analysis was done by the SITECON program [4].

Results: As a result, the classification of the bacterial proteins containing the YfiA homologous domain was made more accurate.

Analysis of the regulatory regions allowed us to identify 8 most statistically significant sites of DNA interaction with regulatory proteins. Comparative analysis of the regulatory regions of the YfiA homologous genes in 55 bacterial genomes, which contain both YfiA and YhbH homologs, demonstrated that 6 of them are determined in other genomes. It is concluded that YfiA may be affected by binding of transcriptional factors of these types.

The work was supported by the Russian Federal Agency for Science and Innovations (contract No. 02.512.11.2165) and the RAS programs “Biosphere origin and evolution” and “Molecular and cellular biology”.

References:

1. Ueta M., Yoshida H., Wada C., Baba T., Mori H., Wada A. (2005) Ribosome binding proteins YhbH and YfiA have opposite functions during 100S formation in the stationary phase of *Escherichia coli*. Genes Cells., 10, 1103-1112.
2. Tikunova N.V., Khlebodarova T.M., Kachko A.V., Stepanenko I.L., Kolchanov N.A. (2007) A computational-experimental approach to designing a polyfunctional genosensor derived from the Escherichia coli gene yfiA promoter, Dokl Biochem Biophys. Nov-Dec;417:357-61.
3. Katoh, Misawa, Kuma and Miyata (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform, Nucleic Acids Res., 30: 3059-3066.
4. SITECON: a tool for detecting conservative conformational and physicochemical properties in transcription factor binding site alignments and for site recognition D. Y. Oshchepkov, E. E. Vityaev, D. A. Grigorovich, E. V. Ignatieva and T. M. Khlebodarova

A MORPHOMECHANICAL APPROACH TO DEVELOPMENT

Beloussov L.V.
Faculty of Biology, Moscow State University, Moscow 119899 Russia
e-mail: morphogenesis@yandex.ru

From a most general point of view, development of an embryo should be regarded as a sequence of space-temporal events leading to progressive and "spontaneous" (as a rule, non-affected from outside) complication (decrease of symmetry order) of its organization at the different structural levels. Hence, embryonic development belongs to the category of self-organizing processes requiring non-linear feedbacks between its dynamic components. Within several last decades, different suggestions on the nature of these feedbacks have been submitted. A morphomechanical approach to development is associated with the idea that the developmental feedbacks are to a great extent associated with mechanical stresses (MS) which are, on one hand, caused by immediately preceded changes of embryonic shape and, on the other hand, determine, according to certain rules, the next set of MS, producing the subsequent shape changes. Contrary to other concepts of developmental control, morphomechanical approach is essentially macroscopic, emphasizing the regulatory role of a large scale geometry and topology of an embryo.

The main arguments supporting morphomechanical approach are the following: in all the studied cases the tissues of developing embryos turned out to be mechanically stressed. MS are arranged according to definite patterns, correlated with the given stage geometry and topology (1); the existence of cells-substrate mechanical feedbacks are both necessary and sufficient for a spatial self-organization of cell populations (2); mechanical factors directly affect gene expression and cell differentiation (3); a cell's ability to produce MS is necessary for its very viability (4).

A general rule is suggested, linking the residual ("passive") MS accumulated in the given cell or embryo part during the immediately preceded period of development with the "active" MS generated later on by this very part. Its central assumption is that the active MS-generated response is directed towards diminishment of the passive MS with a substantial overshoot. Several sets of experimental and modeling data will be presented, showing that this rule is enough robust and can reproduce a number of important morphogenetical events. Also, rather prolonged series of successive events can be presented as a chain of such mutually linked overshooting reactions.

References:

1. Beloussov, L.V., Grabovsky V.I.(2007) Information about a form. *BioSystems* **87**: 204-214.
2. Harris, A.K., Stopak D., Warner P. (1984) Generation of spatially periodic patterns by a mechanical instability: a mechanical alternative to the Turing model.*J. Embryol. Exp. Morphol.* **80**: 1-20.
3. Engler, A.J., Shamik Sen, H. Lee Sweeney and D. E. Discher (2006). Matrix elasticity directs stem lineage specification. *Cell* **126**: 677-689.
4. Chen, Ch.S., Mrksich, M., Huang, S., Whitesides, G.M., Donald E. Ingber, D.E. (1997). Geometric control of cell life and death. Science 276: 1425-1428.

COMPUTATIONAL PIPELINE FOR SMALL RNA DISCOVERY AND EXPRESSION PROFILING BY NEXT GENERATION SEQUENCING

Berezikov E.*, Cuppen E.
Hubrecht Institute, Utrecht, The Netherlands; InteRNA Genomics B.V., Bilthoven, The Netherlands.
e-mail: e.berezikov@niob.knaw.nl
* Corresponding author

Motivation and Aim: Advances in next generation sequencing technology have boosted the area of small RNA research. Massively parallel sequencing of the small RNA complement of cells, tissues and patient samples provides an unbiased approach for the simultaneous discovery and detection of novel and known small RNAs including microRNAs [1]. Although experimental deep sequencing approaches are maturing fast, analysis of the resulting data requires advanced algorithms, state-of-the-art infrastructure and skilled bioinformaticians. Informatics thus becomes a bottleneck for many research laboratories in using deep sequencing technologies.

Methods and Algorithms: We have developed a modular pipeline for the analysis and interpretation of high-throughput small RNA sequencing data [2-5]. Output formats include for example summary tables and text files, fasta sequence files, frequency and heat maps, homology information, tracks on genome browsers, etc.

Results and Conclusions: Reanalysis of publicly available *C. elegans* and *Drosophila* deep-sequencing small RNA datasets using our computational pipeline revealed several highly-confident novel miRNA missed by previous analyses. We also applied the pipeline for comparative analysis of small RNAs from 4 nematode and 4 fish species, and discovered numerous conserved as well as species-specific miRNAs.

Availability: the analysis pipeline is provided as a commercial service by InteRNA Genomics B.V. (http://www.interna-genomics.com).

References:

1. Berezikov et al. (2006) Approaches to microRNA discovery, *Nature Genetics*, **38:** S2-7.
2. Berezikov et al. (2005) Phylogenetic shadowing and computational identification of human microRNA genes, *Cell,* **120:** 21-24.
3. Berezikov et al. (2006) Diversity of microRNAs in human and chimpanzee brain, *Nature Genetics,* **38:** 1375-1377.
4. Houwing et al. (2007) A role for Piwi and piRNAs in germ cell maintenance and transposon silencing in Zebrafish, *Cell*, **129:** 69-82.
5. Berezikov et al. (2007) Mammalian mirtron genes, *Molecular Cell*, **28:** 328-336.

MODELING EVOLUTION OF GENETIC REGULATION IN ARTIFICIAL ORGANISMS

Beslon G.*[1], Sanchez-Dehesa Y.[1], Peña J.-M.[1, 2]
[1] LIRIS Laboratory, INSA-Lyon, France
[2] DATSI Universidad Politecnica de Madrid, Spain
* Corresponding author: e-mail: guillaume.beslon@liris.cnrs.fr

Motivation and Aim: Prokaryote organisms are very diverse, living in different hosts or places and developing very different tasks. Some bacteria, for example, can be also found for example, surviving in acid elements without presence of oxygen, in symbiosis with other organisms (ex. *Buchnera aphidicola*). Bacteria are good example of organism adaptation, being able to adapt themselves to different environments (sometimes "hostile") and to adopt their behavior to new conditions. They have developed different adaptation strategies to evolve and to be adaptable to new conditions, depending on the characteristics of their environment.
Yet, bacterial adaptation strategies can be very diverse, ranging from stochastic fluctuations (e.g., bet-hedging) to genetic regulation. For example, if changes in the environment are slow enough bacteria may mutate and adapt to new conditions, but if they are often submitted to stress, they must develop different adaptation systems such as genetic regulation or noise enhancement.
It is still a matter of research to link adaptation strategies with environment characteristics. In particular, the question of the evolution of genetic networks is an open question. Yet, such a question is very difficult to address experimentally, either because it requires long and complex experimental setups (e.g., experimental evolution) and because results are difficult to analyze given the few knowledge at our disposal. Traceability of changes in genome and identification of selected individual are difficult in experimental evolution. A possible solution is to develop a biological model "in silico" following the main biological foundations of genotype-phenotype mapping and evolution. Such models have already shown their interest to study how bacteria may adapt their genome length to environmental conditions (e.g., bottlenecks) or to their mutation rates.

Methods and Algorithms: To study evolution and adaptation strategies, we have developed an integrated model that includes a realistic genotype-phenotype mapping and that enables the evolution of a genetic network *inside* such a system. In this model, the characteristics of the organisms may change during the evolutionary process (e.g., the number of genes and the genes function can change, so does the connectivity of the regulation network). Moreover, it respects the mains biological organization levels from genome to organism's phenotype and it is compliant with some minimal characteristics: differential genes activation levels, gene positive or negative self-regulation, variable levels of RNA translation and protein production... Finally, organisms are selected on the basis of their phenotype which directly depends on the concentration of the different proteins transcribed from the genome.

Results This model can then be used to design and conduct in silico experimental evolution: populations of individuals evolves in variable environments. Thus, by changing the characteristic times and levels of fluctuation of the environment, we are able to test the evolutionary response of the organism and check when they develop a complex regulation network or, on the opposite, when they evolve toward very simple forms of regulation similar to what can observed e.g., in endosymbiotic bacteria.

Conclusion: The definition of a model of prokaryotic organisms that includes both the evolutionary adaptation mechanisms and genetic regulation mechanisms would be useful to tackle many open questions in the literature, such as: How organisms adapt to environmental changes? In which conditions do genetic regulation arise and what are the rules that govern complexification of regulation networks? How such networks grow during evolution? This proposed model provides the theoretical mechanisms to study the inclusion of new nodes in already existing regulatory networks, and the analysis of the development of new regulatory networks. These issues could help us to answer some of these questions thus providing us a better understanding on prokaryotes evolution.

Availability: RAEVOL software is currently in alpha version, supported by LIRIS group (INSA-Lyon). For any request, please contact directly with the authors.

ANALYSIS OF A LIGHT ENTRAINMENT ON THE MATHEMATICAL MODEL OF MAMMALIAN CIRCADIAN OSCILLATOR

Bezmaternykh K.D.[1, 2*], Podkolodnaya O.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: bezmate@bionet.nsc.ru
* Corresponding author

Motivation and Aim: Internal biological clocks with a period of ~ 24 h (circadian) exist in most organisms from cyanobacteries to mammals and play a key role in processes of sleep-wake cycles; metabolism; stress and growth hormones synthesis, immunity etc. Due to the fact that circadian clocks are capable for entrainment by external signals as light, hormones, physical activity living organisms can set the most energy-wise advantageous behavior. Disorders in circadian clock system cause series of pathologies that can be adjust by light influence in certain time of circadian period. Studying the molecular-genetics regulation of circadian rhythms is therefore important fundamental and applied task.

Methods and Algorithms: For research functional patterns of genetics system that regulates circadian clocks, mathematical modeling and computer analysis were used. As a base of new mathematical model we used Forger-Peskin model of the mammalian circadian clock [3]. Generalized chemical kinetic method [1] and the MGSModeller program [2] were used for extended model development and calculations.

Results: The extended mathematical model of mammalian circadian oscillator was build. Descriptions of genetic subsystems controlled by transcription factors Roralpha and CLOCK/BMAL1 were included. Extended model also contains descriptions of inhibitory effect of transcription factor Rev-Erbalpha on *Bmal1* gene transcription and its auto inhibition. Model's parameters were estimated to fit known experimental data. Model showed adaptation to various phase-shift of light-dark cycle. The rate of adaptation to phase-shift of light-dark cycles was calculated with varying length of free-running period.

References:

1. V.A. Likhoshvai et al. (2001). Generalized chemokinetic method for gene network simulation, *Mol. Biol.*, **35:**1072-1079.
2. F.V. Kazatcev et al. (2008). MGSmodeller'2008 – a computer system for reconstruction, calculation and analysis mathematical models of molecular genetic system . *This issue.*
3. D.B. Forger, C.S. Peskin. (2003). A detailed predictive model of the mammalian circadian clock. *PNAS*, 9;100(25):14806-11.

USING SVM AND A MEASURE OF MOTIF 'SURPRISE' TO DISTINGUISH REGULATORY DNA

Boekhorst R.[1*], Abnizova I.[2], Naumenko F.[2], Wernisch L.[3]
[1] University of Hertfordshire - Dept of Computer Science College Lane, Hatfield - UK
Corresponding author, r.teboekhorst@herts.ac.uk
[2] Welcome Trust Sanger Institute, Hinxton, UK
[3] MRC Biostatistics Unit Institute of Public Health, Cambridge - UK

Motivation and Aim: There are still no satisfactory computational methods to reliably recognize regulatory DNA. Assuming that the main biological and statistical 'signature' of regulatory regions is the presence of multiple regulatory motifs, we are interested in methods that identify those that contribute significantly to the separation of coding, regulatory and non-coding non-regulatory DNA.

Methods and Algorithms: We want to find solutions to this problem by considering both a newly developed feature set and exploring the application of various machine learning classification techniques. In addition, we use two unsupervised pattern recognizing techniques, hierarchical cluster analysis and Principal Component Analysis (PCA), to back up the performance and visualize the results of the supervised SVM.

The novelty of our approach is the use of a new feature representation of DNA sequences as the input for pattern recognition methods. We represent a sequence as a 4^k – dimensional vector, of which the elements are the Z-scores for each possible k-mer with m mismatches of that sequence. The Z-scores measure how likely each k-mer is in comparison to a model assuming nucleotide independence. We then subject the feature set to a hierarchical test procedure that first distinguishes coding from non coding sequences, and in a next step separates regulatory regions from non coding-non regulatory DNA.

Results: We found out that our method captures the difference between functional DNA sequences better than existing related methods, such as those based on string kernels and mismatch kernels. The motifs responsible for clear regulatory DNA separation occurred to be biologically important fragments of known TFBS. We stress the up till now overlooked importance of underrepresented motifs.

Conclusion and Availability: We have shown that a SVM working on sequences representation by Z-scores can be helpful in separating DNA types. Interestingly, it seems that not only over-represented, but also under-represented tri-mers significantly contribute to this separation. The interactive program, written in C sharp, to compute the kernel is available from http://www.mrc-bsu.cam.ac.uk/personal/irina/ for free downloading.

PREDICTION OF PROTEIN ALLERGENICITY BASED ON PROTEIN 3D STRUCTURE PROPERTIES

Bragin A.O.*, Yarkova E.E., Demenkov P.S., Ivanisenko V.A.
Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
Novosibirsk State University, Novosibirsk, Russia
Sobolev Institute of Mathematics, SB RAS, Novosibirsk, Russia
e-mail:
* Corresponding author

Motivation and Aim: Prediction of protein allergenicity is important for many applications in biomedicine and biotechnology. Modern methods for allergenicity prediction are mainly based on homology search in protein primary structure or patterns recognition. Those methods have not high accuracy values.

The aim of this work was to develop the method for prediction protein allergenicity considering not only protein primary structure but also tertiary structure features.

Methods and Algorithms: Conformational peptides are patches on protein surface which satisfied the following conditions: a) the distance between C-alpha atoms of amino acid residues in a conformational peptide should not exceed distance between C-alpha atoms of covalently bound amino acids; b) the amino acid number in conformational peptide was taken 8; c) solvent accessibility of a residue should not be less than 10%.

To create the database of conformational peptides which could be probable allergenic epitopes the SDAP database containing information about known allergenic proteins was used. 3D structure of allergenic proteins was predicted using ESyPred3D and Swiss-Model programs.

The prediction of protein allergenicity was based on search of coincidences between analyzed protein conformational peptides and those from the database.

Results: The database of conformational peptides contains about 70 000 entries for more than 180 allergenic proteins.

The method for prediction of protein allergenicity based on protein 3D structure features was developed. The accuracy of the method was estimated at 85%. The developed method can be widely used for protein allergenicity prediction.

Availability: The database and develop methods will soon be installed on ICG server

Work was supported in part by RFBR: 08-04-91313-IND_a, state contract FASI №02.514.11.4065, interdisciplinary integrative project for basic research of the SB RAS № 115 and RAS presidium program “Molecular and cellular biology”, the grant “Systems biology: computer and experimental approaches.

QUANTUM-CHEMICAL ANALYSIS OF ZN^{2+} BINDING IN WILD-TYPE AND G245C MUTANT OF P53 PROTEIN

Bugakov I.V.[1], Fomin E.S.[2*], Ivanisenko V.A.[2]
[1] Novosibirsk State University, Novosibirsk, Russia
[2] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
e-mail: fomin@bionet.nsc.ru
* Corresponding author

Motivation and Aim: The hypothesis about functional significance of an additional Zn^{2+} binding site, immediately neighboring the normal zinc-binding site, was confirmed by our molecular simulation calculations [1]. It was shown that the interaction energies of a Zn^{2+} ion with both normal and additional sites of p53 G245C mutant were comparable, with the normal site being energetically handicapped. However, former energy calculations could provide only a qualitative explanation of the impaired function of the p53 G245C mutant. The quantitative evaluation of energy differences demands quantum-mechanical methods because of a multicenter donor-acceptor character of the bond between Zn, S, and N atom, which cannot be modeled in the force field approximation.

Methods and Algorithms: The bonding energies of Zn^{2+} in both zinc-binding sites of the mutant G245C form were calculated using GAMESS software package with B3LYP Density Functional Theory method on the 6-31G(d,p) basis. The spatial structures of the human p53 core domain were used. They included the wild-type structure (PDB ID 1gzh) and the model structure of the G245C mutant [2]. The initial position of Zn^{2+} in the G245C mutant was defined using the PDBSiteScan program [3]. The positions of all atoms in both binding sites were optimized by the MOLKERN software package [4].

As molecules containing thousands atoms are beyond the reach of quantum-mechanical calculations, some small (no more than 70 atoms) clusters were modeled. Such clusters involved all the atoms of the binding site itself and atoms of the nearest coordinate environment. All the dangling bonds were protonated. In such cluster modeling it was presumed that the distant environment distortion had an indirect influence upon Zn^{2+} binding energy shifting through local environment configuration changes.

Conclusion: The calculation results are in qualitative agreement with earlier results [1] and demonstrate the importance of taking into account the exchange-correlated interaction energy changes for Zn^{2+} binding.

References:

1. E.S.Fomin, V.A.Ivanisenko (2008) Corroboration of the Functional Role of the Additional Zinc Binding Site in the G245 Mutant Form of the p53 Protein, *Biofizika*, in press.
2. V.A.Ivanisenko, S.S.Pintus, D.A.Grigorovich, N.A.Kolchanov (2005) *Nucleic Acids Res*., **33**: D183.
3. V.A.Ivanisenko, S.S.Pintus, D.A.Grigorovich, N.A.Kolchanov (2004) *Nucleic Acids Res.*, **32**:W549.
4. E.S.Fomin, N.A.Alemasov, A.S.Chirtsov, A.E.Fomin (2007) MOLKERN as new effective engine for drug discovery software 4th International Symposium Computational Methods in Toxicology and Pharmacology Integrating Internet Resources (CMTPI-2007): 97.

PHYLOGENETIC ANALYSIS OF NEURALIZED GENES AND PROTEINS

Bukharina T.A.*[1], Gunbin K.V.[1], Furman D.P.[1, 2]
[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
[2] Novosibirsk State University, Novosibirsk, Russia
e-mail: bukharina@bionet.nsc.ru
* Corresponding author

Motivation and Aim: The neurogenic gene *neuralized* plays a key role in determining the cell development pathway during the formation of the neuroectoderm. Homologs of *neur* gene play analogous role in other biological species. It is known that *neur* genes can also control other biological processes. In particular, a human homolog of the Drosophila *neuralized* gene has been described as a potential tumor suppressor gene in malignant astrocytomas. NEUR protein contains two large repeats termed Neuralized Homology Repeats (NHRs), which are responsible for protein–protein interactions, and a RING finger (RF) domain, associated with an ubiquitin ligase activity of this protein. The goal of this work was to study the evolutionary modes for *neuralized* genes and the corresponding proteins by analyzing the reconstructed phylogenetic trees.

Methods and Algorithms: Homologous proteins and nucleotide sequences were extracted from the GenBank and Ensemble databases. Sequences were aligned using the PROMALS program [1]. Phylogenetic trees and ancestral sequences were reconstructed using PhyML-aLRT 1.1 [2] and FastML 2.02 [3]. The adaptive evolutionary events were searched for by the computer test described in [4] and rapidly evolving protein regions, by the Rate4Site 2.01 program [5].

Results and Conclusion: Computer analysis of the sequences and phylogenetic trees for the *neuralized* genes and the corresponding proteins has detected a correlation between the adaptive evolutionary events and emergence of large taxa, in particular, Vertebrata. It has been demonstrated that NHR1 is a rapidly evolving region in the protein analyzed.

Acknowledgements: The work was supported by the Russian Academy of Sciences (RAS) program no. 2 for basic studied *Molecular and Cell Biology* (project no. 10.4) and the RAS Presidium program *Biosphere Origin and Evolution* (integrated project of the Siberian Branch of RAS no. 18.13).

References:

1. J. Pei, N.V. Grishin (2007) PROMALS: towards accurate multiple sequence alignments of distantly related proteins. *Bioinformatics,* **23**: 802-808.
2. M. Anisimova, O. Gascuel (2006) Approximate likelihood ratio test for branches: A fast, accurate and powerful alternative. *Syst. Biol.*, **55**: 539-552.
3. T. Pupko et al. (2002) A branch-and-bound algorithm for the inference of ancestral amino-acid sequences when the replacement rate varies among sites: Application to the evolution of five gene families. *Bioinformatics,* **18**: 1116-1123.
4. K.V. Gunbin et al. (2007) The evolution of the Hh-signaling pathway genes: a computer-assisted study, *In Silico Biol*., **7**: 333-354.
5. I. Mayrose et al. (2004) Comparison of site-specific rate-inference methods: Bayesian methods are superior. *Mol. Biol. Evol.*, **21**: 1781-1791.

DATABASE NEUROGENESIS ON BRISTLE PATTERN FORMATION IN *D. MELANOGASTER*

Bukharina T.A.*[1], Furman D.P.[1, 2]
[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
[2] Novosibirsk State University, Novosibirsk, Russia
e-mail: bukharina@bionet.nsc.ru
* Corresponding author

Motivation and Aim: Macrochaetae (bristles) are the external sensory organs located at stringently determined positions on the fly's body. The morphogenesis of macrochaetae consists of three stages: (1) separation of proneural clusters, groups of 20–30 cells, each possessing the potential of a sensory organ precursor (SOP) cell, from the cells of wing imaginal disc; (2) positioning of SOP cell within a proneural cluster; and (3) two asymmetric divisions of SOP cell and differentiation of the daughter cells into the definitive mechanoreceptor components—trichogen, tormogen, neuron, and thecogen. The bristle formation has complex molecular genetic control [1]. Integrated description and systems analysis of functioning of such systems require oriented databases and computer technologies.

Methods and Algorithms: GeneNet technology [2] was used for reconstruction of the gene networks.

Results: Based on annotation of over 600 papers, the database Neurogenesis was developed; it contains information about 217 components (80 genes, 112 proteins and protein complexes, 6 RNAs, and 16 processes) and 289 links between them. The gene networks "Neurogenesis (determination)" and "Neurogenesis (asymmetric division)", describing stages 1–2 and 3, respectively, were reconstructed.

Conclusion: Logical analysis of the gene network "Neurogenesis (determination)" demonstrates that the content of the products of *achaete–scute (AS-C)* gene complex, reaching its maximum in the SOP cell, is the critical factor determining the neural developmental pathway. The key components of this network are *AS-C* and the mechanisms providing the necessary concentration of AS-C protein in the proneural cluster cells and SOP cell, namely, self-regulation of *AS-C* genes (activation by AS-C/DA heterodimers and repression by AS-C/EMC heterodimers), trans-regulation of *AS-C* genes (activation by the transcription factors CHARLATAN and SENSELESS and repression by the transcription factor HAIRY), and EGFR and Notch signaling pathways.

The key elements of the gene network "Neurogenesis (asymmetric division)" are the NUMB and NEUR proteins, determining specialization of the daughter cells; the mechanisms providing an asymmetric distribution of these proteins in the SOP cell and an asymmetric segregation between the daughter cells via formation of the protein complexes aPKC, PAR, BAZOOKA, PINS, DLG, and LGL; and the mechanisms determining a correct mutual orientation of the protein determinants and cell division plane with involvement of the G protein signaling system.

Availability:
http://wwwmgs.bionet.nsc.ru/mgs/gnw/genenet/viewer/Neurogenesys_determination.html

Acknowledgements: The work was supported by the Russian Academy of Sciences (RAS) program no. 2 for basic studied *Molecular and Cell Biology* (project no. 10.4) and the RAS Presidium program *Biosphere Origin and Evolution* (integrated project of the Siberian Branch of RAS no. 18.13).

STUDYING OF THE QUESTION ABOUT NECESSARY NUMBER OF DNA SEQUENCES TO CARRY OUT OF POPULATION-GENETIC RESEARCHES WITH THE HELP OF IMITATING COMPUTER MODEL

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

Motivation and Aim: In that time molecular data are widely applied to carrying out of population-genetic researches. In particular, with the help of special statistical methods on the basis of distinctions of pair's differences in DNA sequences it is possible to estimate the effective size of populations and intensity of genetic flows between various populations. All these methods are based on the developed theory inter population genetic flows and the theory of genetic drift.

In connection with that process of sequencing of DNA still remains expensive, and the material for the genetic analysis is difficultly accessible, there is a question about enough number DNA sequences to obtain statistically reliable result.

Methods and Algorithms: To determine the number of sequences for carrying out of the genetic analysis, it is necessary to investigate such populations in which considered fragment of DNA is known for all organisms. The necessary population parameter is originally counted up proceeding from all known sequences. Further we take the limited number of sequences and calculate the population parameters used this limited data set. The received result is with result for all population. Consistently increasing the limited sample of DNA and comparing result with full population sample, it is possible to estimate convergence of result of the genetic analysis and determine the optimal number of DNA sequences necessary to carrying out of researches.

To make the described experiment on populations of real organisms it is practically impossible. However processes of genetic drift and genetic flows can be modeled easily with the help of computers. Using computer models, we can receive full sets of DNA sequences of populations of "computer" organisms after various parameters (the effective size of a population, probability of a mutation per nucleotide per generation, intensity of an exchange of a genetic material).

*Results, Conclusion and Availability:*In our work we used computer models for research of convergence of the F_{st} criterion that used to estimate of genetic flows, and quantity θ that used for an estimation of the effective size of populations ($\theta=4N_e\lambda$, where N_e is an effective size of a population, and λ - probability of a mutation per nucleotide per generation).

In our research we establish that at the various effective sizes of a population and various probability of a mutation and at number of sequences is 30 the mistake of calculation F_{st} and θ is about 25 %. To obtain more reliable result we mast take about 50-60 samples of DNA sequences from researched population of organisms.

«KARYOSTATANALYSIS» - A SOFTWARE FOR CHROMOSOMAL SETS MORPHOMETRIC ANALYSIS

Bukin Yu.S.*, Natyaganova A.V.
Limnological Institute of RAS SB, Irkutsk, 664033, Russia
e-mail: bukinyura@mail.ru, avn@lin.irk.ru

Motivation and Aim: One of first stages of organisms karyological studies is description of chromosomal sets (karyotyping). This procedure includes morphometric analysis due to which an averaged scheme (karyogram) of morphological structure for species chromosomal set is created. This is a karyotype. As a rule, chromosomes are measured on images of methaphase plates obtained from 10 individuals. On the base of data obtained, several morphometric parameters are calculated. Their calculation even using a well-known software Exel takes rather long time.

Methods and Algorithms: We have developed a software for PC which simplifies the statistical treatment of data obtained by morphometric analysis of organisms chromosomal sets. This ware calculates practically in a moment the following parameters necessary for the typical procedure of karyotype description: average values (as well as average errors) of limbs lengths, of relative lengths of chromosomal pairs and of centromeric indices values. Besides, the software allows to calculate the variability factor of chromosomal pairs relative lengths and index of karyotype asymmetry, these are indicators which contain much information for comparative karyotypic analysis. Calculation algorithms of the parameters indicated are done in computer language C++.

Results: The calculated values of those parameters can be converted into other software программы (Word, Exel, Statistika) and presented as tables and karyograms.

Availability: Hypotetic name of this software is «Karyostatanalysis». It is planned to present this software on the web-site of the Limnological Institute of RAS SB at the end of 2008.

Conclusion: The software proposed can be of interest for people working in the field of karyology and recommended for cytogenetic practice at biological departments of high schools.

This work is supported by RFBR, grants NoNo 04- 04-48945-a, 07-04-01410-a.

GENETIC LINKAGE ANALYSIS CHALLENGES ON A DISTRIBUTED GRID ENVIRONMENT

Calabria A., Pasquale D., Orro A., Trombetti G., Milanesi L.*
Institute of Biomedical Technologies ITB-CNR, Milano, Italy
e-mail: luciano.milanesi@itb.cnr.it
*Corresponding author

Motivation and Aim: Linkage Analysis is a statistical method for detecting genetic linkage between disease loci and markers of known locations by following their inheritance in families through the generations. It is a NP-hard problem and the computational cost and memory requirements of the major algorithms proposed [2-4] grows exponentially with pedigree size and markers' number. Implementations of the mentioned algorithms reflect these limits making analyses of medium/large data sets very hard on a single CPU. The aim of the present work is to speed up the execution of linkage analysis challenges, that is the execution of linkage workflow on large data sets. The challenge is launched on the EGEE Grid [1] infrastructure, distributing the needed processes among different computing elements. In this context, a user friendly web based access to grid resources has been implemented.

Methods and Algorithms: Many of the most used linkage analysis software have been ported into the Grid environment and a suitable workflow has been designed and implemented in order to exploit the parallelism of the workflow from the input to the output by distributing each task on different computing elements. At a higher level, all these steps are parameterized and monitored with the support of a web interface, designed to simplify and speed up the whole process of linkage analysis. At a lower level, the workflow progress is supported by a reliable, scalable and secure grid system facility, called VNAS [5], which monitors each single grid process and ensures its elaboration success managing the resubmission of failed jobs automatically.

Results: Test results show that when reaching the computational limits in data set size there is a real benefit in the use of our implementation. We experienced an improvement of about 65% in computational time compared to a desktop computer execution when adopting Genehunter [6] with 16 individuals and 100 markers, while increasing the number of individuals some elaborations, still successfully working on the Grid, resulted infeasible on a desktop PC due to memory overflow.

Conclusion: We developed a system which enables the user to launch genetic linkage analysis calculations for medium to large challenges over the grid infrastructure and we designed a suitable workflow to achieve parallel processing of the tasks: a simple and complete web user interface provides an easier approach to linkage analysis software on the grid environment, while lower-level reliable software manages grid interactions. However, a possible slowdown of the grid infrastructure may lead to a reduction in the efficiency and reliability of the proposed pipeline compared to single workstation. The approach used has been designed to work within the EGEE framework, granting web access to GRID technology with very basic informatics knowledge requirements.

References:

1. F. Gagliardi, B. Jones, F. Grey, M.E. Begin, M. Heikkurinen (2005) Building aninfrastructure for scientific Grid computing: status and goals of the EGEE project, *Philosophical Transactions: Mathematical, Physical and Engineering Sciences* **363**:1729-1742
2. R.C.Elston, J.Stewart (1971) A general model for the analysis of pedigree data, *Hum Hered*, **21**: 523-542.
3. E.S.Lander, P.Green (1987) Construction of multilocus genetic linkage maps in humans, *Proc. Nati. Acad. Sci.*, **84**: 2363-2367.
4. A.A.Schaffer (1996) Faster linkage analysis computations for pedigrees with loops or unused alleles, *Hum Hered.,* **46(4)**: 226-35.
5. G.A.Trombetti et al. (2007) Data handling strategies for high throughput pyrosequencers, BMC Bioinformatics, **8(1)**:S22.
6. L. Kruglyak, M.J. Daly, M.P. Reeve-Daly, E.S. Lander (1996) Parametric and nonparametric linkage analysis: a unified multipoint approach, *Am J Hum Genet,* **58(6)**:1347-63.

ANNOTATION OF LUNG-SCREENING IMAGES AND 2D-E PROTEOMIC ANALYSIS FOR EARLY DIAGNOSIS OF LUNG CANCER THROUGH FEDERATED BIOBANKS

Cataldo R.*[1], Quarta M.[2], Agrusti A.[1], Nunzio G.[1], Maglio S.[1], Fantacci M.E.[3], Bagagli F.[3], Favetta M.[4], Massafra A.[4], Mercurio G.*[5]

[1]Department of Science of Materials – University of Salento- Lecce and INFN Lecce- Italy
e-mail: rosella.cataldo@unile.it
[2]Department of Mathematics – University of Salento- Lecce and INFN Lecce – Italy
[3] Department of Physics – University of Pisa and INFN Pisa – Italy
[4] Department of Physics – University of Salento- Lecce and INFN Lecce – Italy
[5]*Istituto Tecnologie Biomediche CNR - Italy*
e-mail: gregorio.mercurio@itb.cnr.it
*Corresponding author

Abstract: Motivation and Aim: Lung cancer is the leading cause of death by tumor in Europe. New possibilities for an early diagnosis of lung cancer are provided by means of Computerized Tomography (CT) as a diagnostic device for small lung nodules early detection; so computerized techniques for the automated analysis of these scans are quickly becoming a practical necessity and are expected to provide valuable assistance to radiologists. Because the goal of an automated lung nodule detection method is to improve human detection performance, an essential component of any automated method is the process through which computer-generated results and information are conveyed to the radiologist: this improvement is possible because of expert radiologists can annotate the scans, using an automated tool and a common annotation protocol, so the results of the radiological examination and the CAD evidences can be compared. Moreover, proteomic analysis studies on plasma were aimed at the early detection of lung cancer through the identification of protein profiles and molecules involved in the pathogenesis, progression and sensitivity to specific therapies.

Methods and Algorithms: Medical Application on a Grid Infrastructure Connection (MAGIC-5) is an italian academic project, involving physicists and physicians, engaged in the effort to develop CAD methods for detection of smaller nodules in screening programs. Focus of our study is the development of a prototype of working environment that can improve flexibility and efficiency in creating, storing, and exchanging annotations of medical images, associated with protein images produced by two-dimensional electrophoresis (2-DE). Therefore, we outline the most important features of the annotation protocol, that comprises both the inclusion criteria in collecting images and the clinical information consistent with our aims.

Results: We describe the technical features of the low-cost automatic annotation tools, developed under Open Source platform, used by radiologists and biologists to annotate scans and protein images. Finally, we make an attempt to relate the annotations of the CT database with the results of the analysis of 2DE images, in order to identify potential pathological biomarkers by same patients.

Conclusion: Our aim is to explore if proteomic analysis of the plasma could increase the sensibility and diagnostic specificity of CT analysis, also through a federation of heterogeneous biobanks of lung-screening and 2-DE images.

MODEL OF PERFECT TANDEM REPEAT WITH RANDOM PATTERN FOR LATENT PERIODICITY RECOGNITION IN BIOLOGICAL SEQUENCES

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. Bauman, Moscow, Russia
e-mail: maramaria@yandex.ru
[*] Corresponding author

Motivation and Aim: The problem of latent periodicity recognition in biological sequences is topical due to its practical application in biology and medicine: genotyping of microorganisms, the revelation of the risk source for hereditary diseases and recognition of functional structures in DNA and proteins. At present, latent periodicity recognition is based on a notion of approximate tandem repeat [1]. The joint use of combinatorial and dynamic programming methods has been received the most propagation in periodicity recognition [1]. However, redundancy and instability of the results have been noted [2,3]. Therefore, statistical methods are applied for estimation of periodicity pattern size. Unreliability of these methods, observed in practice, is due to both limited statistical material and orientation of the estimation search at maximum heterogeneity manifestation in sequence analyzed [3]. The present work is aimed at elaboration of the methods raising reliability and stability of pattern periodicity recognition.

Methods and Algorithms: To solve the problem put by an original spectral-statistical approach has been proposed [3]. A stochastic model of perfect tandem repeat with random periodicity pattern, introduced in [3], has served as substantiation of the approach, applied in practical conditions. This model adequately describes heterogeneity manifestation in the text strings.

Results: Efficiency of spectral-statistical approach has been proved on TRDB database of approximate tandem repeats (http://tandem.bu.edu/cgi-bin/trdb/trdb.exe). The approach allows to receive stable and reliable estimation of periodicity pattern size. It has been shown [3], such an estimation optimizes a process of pattern search and in a number of cases allows to surpass the results obtained by generally recognized Tandem Repeats Finder [1].

Conclusion: The proposed spectral-statistical approach is effective for recognition of periodicity pattern in approximate tandem repeats. The considered model of perfect tandem repeat with random pattern allows us to introduce a new notion of latent periodicity for biological sequences, which expands notion of approximate tandem repeat.

References:

1. G.Benson (1999) Tandem repeats finder: a program to analyze DNA sequences, *Nucl. Acids Res.*, **27**: 573-580.
2. V.Boeva et al. (2006) Short fuzzy tandem repeats in genomic sequences, identification, and possible role in regulation of gene expression, *Bioinformatics*, **22**: 676-684.
3. M.Chaley, V.Kutyrkin (2008) Model of perfect tandem repeat with random pattern and empirical homogeneity testing poly-criteria for latent periodicity revelation in biological sequences, *Mathematical Biosciences*, **211**: 186-204.

THE WITHIN-INDIVIDUAL BASIS OF BETWEEN-INDIVIDUAL DIFFERENCES

Cherdantsev V.G., Scobeyeva V.A.*
Department of Biological Evolution, Faculty of Biology, MSU, Moscow, Russia
e-mail: skobei-khanum@yandex.ru
* Corresponding author

Motivation and Aim: We extend the notion of repeatability borrowed from the quantitative genetics to the developing systems in order to show that individual morphological variation bases on within-individual differences whose spatial unfolding ("repeatability") is inherent to directional change in spatiotemporal geometry of developing embryos.

Methods and Algorithms: We studied spatiotemporal dynamics of metric morphological characters in amphibian and teleost gastrulation both in living embryos and samples of synchronously fixed embryos of the same age. The embryos were genetically homogeneous (siblings) and developed in optimal environmental conditions with no embryonic mortality. Routine methods of multivariate statistical analysis have proved to be sufficient for reconstructing of both the variation patterns and geometric algorithms of robust morphological transformations.

Results: When in a sample of the adult organisms the amount of variance in metric characters expressed in the coefficients of variation (*CV*) values is about 20% and more the biologist would suspect that something is wrong with sampling. Meanwhile, when we were considering metric characters of the embryonic structures subject to shaping, the values of the *CV* at a level of 15-20% corresponded not to the upper, but rather to the lower limit of variation. Variation of this order of magnitude is common to siblings developing in optimal conditions and having the same developmental age. It follows that the normal developmental variation can be referred neither to genetic nor environmental components of variance. It merits being considered as a special developmental component of natural phenotypic variance – the inherent developmental variation (IDV). In contrast to so-called "intangible" developmental variation that accumulates with time, the IDV increased with initiating of each new morphogenetic shaping and decreased, as the developing structure acquired its eventual shape. The IDV estimated at the individual pathways of amphibian and teleost embryos during gastrulation was of the same order of magnitude that variance estimated in synchronous samples of the embryos. The behavior of individual pathways is close to deterministic chaos, which means individual pathways of the embryonic shaping being instable in spite of deterministic trends of the shaping itself. It follows that between-individual differences are reproducible, if, and only if, they have a within-individual analog.

Conclusion: In general, morphogenesis is subject to the spatial unfolding principle, which means the spatial series of differently shaped structures being equivalent to the successions of shaping of the same structure. From the analysis of quantitative morphological data, we infer a general law that, in the spatial unfolding of shaping of the embryonic regions, the proportion of within- to between-individual variation in a given region depends only on its spatiotemporal position in the unfolding. The later ("younger") regions of the developing structure prove to be more variable and develop faster than the earlier ("older") ones. This holds both for ordinary series of repeatable structures and for consecutive developmental stages of single embryonic structures, such as the dorsal blastopore lip in amphibians and embryonic shield in teleosts. The system is parametric, in the sense that the shape of "older" regions affords parameters for the development of "younger" regions, and dynamical, in the sense that the shape of the whole unfolding of differently aged regions behaves like a single dynamical system. The general principle is, just like in Fisher's theorem of natural selection, that the system moves towards minimization of spatial variance of the whole unfolding with a speed equal to the amount of variance in a given region.

BIOINFORMATIONAL MODELS FOR TESTING OF MEDICINAL PLANTS

Cherkashin A.K., Popov P.L.
V.B. Sochava Institute of Geography, SB RAS, Irkutsk, Russia
e-mail: cherk@mail.icc.ru

Motivation and Aim: The existence of the non-genetic heredity and causality in non-living nature, and in superbiological societal and geographical structures makes it possible to regard bioinformatics as a part of general informatics and pose the question as to the existence of objective reasons for formation of the typology and classification of subjects in which the genetic heredity is merely an intermediate link between subject entities and observed phenomena.

Methods and Algorithms: When modeling the space of an entity ((bio)information space) we proceed from the following statements: 1) all reality consists, as a minimum, from three isolated layers: the information layer, the energy layer, and the material (observed) layer; 2) all objects exist simultaneously in all layers connected informationally via mappings; and 3) the structure of the information space (environment) is represented by a multispiral or by an equivalent fractal hierarchical structure of a classification of the entities of objects; 4) the classification position is encoded in the interval [0,1] by a multiplace number; 5) the digits of the code are quantum numbers of the solution for a differential equation of the information environment correlating to each code the eigen-functions of fluctuation and corresponding to energy; 6) any material phenomenon is connected with a definite entity (the code of classification position) through the energy function (EF), i.e. the logarithm of realization reliability of observed processes or phenomena; 7) the EF ultimately determines the intensity and directedness of observed processes and can be estimated quantitatively from observation results. The code has the form 0,*a*, where the finite sequence *a*=*a*1*a*2*a*3,…, and each digit *ai* reflects the position of the image of the object on the coils of the multispiral or at hierarchical levels of classification with a decrease in the scale.

Any observed phenomenon is a coincidence of circumstances that are individually encoded by fragments of the multispiral in the form of a vector *a* in a bioinformation space. The list of circumstances in the request structure is formalized in the form of a set of such vectors, the scalar product of which in the form of a bilinear function provides the answer to the posed questions in a quantitative form.

Results: Such an algorithm for processing bioinformation is realized in solving the problems of testing medicinal remedies of vegetation origin to be used to treat viral diseases. A list of 674 flora species was drawn up for the territory of the USSR, which were used to treat 18 viral infections of humans and animals. Groups of plants offering promise for treatment of viral diseases were identified through a classification of the plants and diseases using the bilinear bioinformation model.

Conclusion: The suggested approach holds promise for the advancement of general bioinformatics, because it is based on identifying the invariant structure of the information space connected uniquely with observed processes at any scale of their manifestation.

UBIDENT: A NEW TOOL FOR MS-BASED IDENTIFICATION OF UBIQUITYLATION SITES

Chernorudskiy A.L.[1*], Astashev M.E.[2], Gainullin M.R.[1]
[1] Nizhny Novgorod State Medical Academy, Nizhny Novgorod, Russia
[2] Institute of Cell Biophysics RAS, Pushchino, Russia
e-mail: chalbio@mail.ru
[*] Corresponding author

Motivation and Aim: Identification of protein post-translational modifications is one of the hottest topics in a modern proteomics. Protein ubiquitylation is of especial interest due to its role in regulation of diverse cellular processes and involvement in pathogenesis of severe human diseases [1]. To facilitate identification of particular ubiquitylation sites in proteins by mass spectrometry (MS), we developed UbIdent, a new tool for calculating precise masses of peptides bearing modification.

Methods and Algorithms: UbIdent is initially developed using Borland Delphi framework. UbIdent searches a protein sequence for proteolysis sites, calculates precise peptide masses and generates a list of modified peptides. At the moment the search area is limited by a single sequence provided by user in a FASTA format, but we are going to enlarge it to a database level soon. Integration of UbIdent into the UbiProt Database (http://ubiprot.org.ru/), containing information about a wide set of ubiquitylated proteins, is currently in progress.

Results: Identification of ubiquitylation sites by MS is currently based on the search for signature peptides. Peptides containing modification do not undergo cleavage after modified lysine residue during trypsinolysis. Besides that, these tryptic peptides have specific ubiquitin-derived diglycine (GG-) tags on modified lysines, with an additional mass of 114.1 Da. However, it was recently demonstrated that ubiquitylation sites could be identified by other types of signature peptides containing LRGG-tag (383.2 Da) on internal lysine residues or canonical GG-tag on the C-terminus of ubiquitylated peptides [2]. UbIdent calculates peptide masses taking into account all 3 possible types of signature peptides. This allows accurate identification of ubiquitylation sites from MS data, giving more results than standard search algorithms.

Conclusion and Availability: UbIdent tool will improve MS-based identification of ubiquitylation sites and therefore aid future studies on ubiquitin system and related fields. UbIdent will be freely available for non-commercial usage after integration into UbiProt Database at http://ubiprot.org.ru/. The software package is currently available on request from the authors.

References:

1. M.H.Glickman, A.Ciechanover (2002) The ubiquitin-proteasome proteolytic pathway: destruction for the sake of construction, *Physiol. Rev.*, **82**: 373-428.
2. N.J.Denis et al. (2007) Tryptic digestion of ubiquitin standards reveals an improved strategy for identifying ubiquitinated proteins by mass spectrometry, *Proteomics*, **7**: 868-874.

COMPARATIVE ANALYSIS OF MOUSE CHROMOSOMAL DNA DIGESTION WITH RESTRICTION ENDONUCLEASES *IN VITRO* AND *IN SILICO*

Chernukhin V.A.*, Abdurashitov M.A., Tomilov V.N., Gonchar D.A., Degtyarev S.Kh.
SibEnzyme Ltd., Ak. Timakova 2/12, Novosibirsk, 630117 Russia,
e-mail: valera@sibenzyme.ru
* Corresponding author

Theoretical diagrams of mouse chromosomal DNA cleavage at 17 nucleotide sequences 4-6 bp in length, which are the recognition sites of restriction endonucleases, have been plotted based on earlier suggested method of mammalian genomes restriction analysis *in silico [1]*. Analysis of mouse LINE 1 repeats, presented in database, and products of these repeats cleavage at the same nucleotide sequences has been carried out. In general, the diagrams of mouse chromosomal DNA digestion correspond to diagrams of LINE1-repeats cleavage. Mouse chromosomal DNA hydrolysis with restriction endonucleases, which possess the corresponding recognition sites, has been performed. A comparison of DNA hydrolysis patterns and the plotted diagrams has revealed a good correspondence between the experimental and theoretical data. Only LINE1 repeats and satellite DNA cleavage products are visualized in experiments on chromosomal DNA cleavage with subsequent gel-electrophoresis. Mouse chromosomal DNA cleavage with new methyl-dependent site-specific DNA endonucleases BlsI, GlaI and GluI has been performed.

References:

1. M.A. Abdurashitov, V.N. Tomilov*, V.A. Chernukhin, D.A. Gonchar, S.Kh. Degtyarev (2006) Mammalian chromosomal DNA digestion with restriction endonucleases *in silico, Ovchinnikov bulletin of biotechnology and physical and chemical biology,* **2:** 29-38

AN ALGORITHM FOR PROTEIN CONFORMATIONAL FLEXIBILITY PREDICTION

Chirtsov A.S.[1], Fomin E.S.[2] *
[1] Novosibirsk State University, Novosibirsk, Russia
[2] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
e-mail: fomin@bionet.nsc.ru
* Corresponding author

Motivation and Aim: The broad diversity of protein functions existing in the nature depends largely on changes and transformations of their spatial structures. The basic method to explore these transformations at the molecular level is molecular dynamics. However, the computational burden of molecular dynamics is very high. The coarse-graining methods provide one of the ways to reduce this computational burden, because they use some large fragments comprising from several to hundreds atoms rather than individual atoms.

The approximate methods of protein conformational flexibility calculation are based on molecule graph representation analysis. Now they are extensively used because they provide good accuracy without significant increase of computational complexity [1]. In these methods, molecules are represented as directional graphs with atoms as vertices and chemical bonds as edges (hydrogen, covalent, etc.). Analysis of the graph topology allows recognizing groups of atoms corresponding to rigid or flexible fragments.

Results: In current work we offer a new method for three-dimensional graph analysis. The method is based on the two-dimensional "The Pebble Game" graph algorithm, which uses Laman's theorem [2]. Although the Laman's theorem is invalid for the general three-dimensional case, fortunately, the number of situations when some errors can appear is insignificant and can be analyzed according to discrepancy between the three-dimensional and two-dimensional cases in some characteristics. The proposed method demonstrated good accuracy with insignificant loss of computational performance. The method is supposed to be used in the program component library MOLKERN, which is being developed at the Institute of Cytology and Genetics SB RAS [3].

References:

1. A.Ahmed, S.Kazemi, H.Gohlke (2007) Protein Flexibility and Mobility in Structure Based Drug Design, *Frontiers in Drug Design & Discovery*, **3**.
2. D.Jacobs, B.Hendrickson (1997) An Algorithm for Two-Dimensional Rigidity Percolation: The Pebble Game, *J. Comp. Phys.*, **137**: 346–365.
3. E.S.Fomin, N.A.Alemasov, A.S.Chirtsov, A.E.Fomin (2007) MOLKERN as new effective engine for drug discovery software, *4th International Symposium Computational Methods in Toxicology and Pharmacology Integrating Internet Resources*: 97.

TRANSMEMBRANE DOMAINS OF PROTEINS AS PHARMACEUTICAL TARGETS: KNOWLEDGE-BASED COMPUTATIONAL STRUCTURE PREDICTION

Chugunov A.O.*, Efremov R.G.
Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Russian Academy of Sciences, ul. Miklukho-Maklaya 16/10, GSP Moscow, 117997 Russia.
* Corresponding author. E-mail: batch2k@yandex.ru.

Motivation and aim: Integral membrane proteins (MPs) represent the core of the cell machinery that is responsible for transmembrane (TM) signaling, perception of various stimuli, transport and other vital functions — therefore making up a considerable part of pharmaceutical targets. TM domains of MPs propagate biological signals into the cell — *via* conformational switch in G–protein-coupled receptors (GPCRs) structure upon ligands binding, or phosphorylation of cytoplasmic domain of receptor tyrosine kinases (RTKs) that follows dimerization, or in some other way. Given the difficulties related to experimental structural characterization of the most novel MPs, computational structure prediction of TM domains of essential MPs may be of substantial help for pharmaceutical industry and structure-based design of new drugs — e.g., high-affine GPCR ligands or antitumor inhibitors of RTK dimerization.

Methods and Algorithms: In consideration of constantly (notwithstanding rather slow) growing body of structural data on TM domains, it's becoming possible to take into account peculiarities of MPs organization while building and optimizing computer models of TM domains. We propose a method designed for evaluation of "packing quality" of TM domains models — e.g., homology-built. The approach is based on determination of structural preferences of aminoacid residues types to predefined "classes of membrane–protein environment" over the representative set of available MPs structures, thus enabling construction of the "membrane score" function, estimating the correctness of the model under evaluation (Chugunov *et al.*, 2007).

Results: The "membrane score" method proved itself useful for identification of close-to-native structures of 7-TM proteins like rhodopsin, bacteriorhodopsins, and TM-dimers like glycophorin A and BNIP3 — among massive sets of native-like and erroneous models. Other capabilities — such as detection of alignment errors, optimal choice of template for modeling, coarse-grained optimization of models — are inherent in the method as well.

Conclusion: Consideration of MPs distinctive features is prerequisite for successful modeling of TM domains structures which may be of help in the drug-design process. The "membrane score" method, especially if accompanied by efficient conformational sampling strategies, will supply industry with realistic models of TM domains in proteins.

Reference:

1. Chugunov A.O., Novoseletsky V.N., Nolde D.E., Arseniev A.S., Efremov R.G. (2007). A method to assess packing quality of transmembrane α-helices in proteins. I. Parameterization using structural data. *J. Chem. Inf. Model.* **47**, 1150–1162 and … II. Validation by "correct *vs.* misleading" test. *J. Chem. Inf. Model.* **47**, 1163–1170.

BIOMOLECULA.RU: POPULARIZATION OF LIFE SCIENCE IN RUSSIA

***Chugunov A.O.*[1]*, Natalin P.B., Polyansky A.A.*[1]**

[1] *Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Russian Academy of Sciences, ul. Miklukho-Maklaya 16/10, GSP Moscow, 117997 Russia.*

* Corresponding author. e-mail: batch2k@yandex.ru

Biology, particularly the parts of it which study living systems at the molecular level, is the one of the most rapidly developing fundamental sciences today — what creates the necessary prerequisites for practical application of scientific knowledge in such socially important areas as biotechnology, molecular medicine, *etc*. Given this, the problem of science popularization for the society is becoming more and more prominent. There is no doubt that taxpayers, government officials and businessmen have to understand how do scientists spend their money and what are the possible benefits the investments can bring. Needless to say that scientific and technological potential of the State defines its position on the international scene. The high authority of science as a social institution which is accepted by the society lays at the basis of such potential.

Western civilization long ago had become aware of the necessity of making science explained to a layman. Nowadays it is the part of a long lasting cultural tradition. Many high profile scientific journals, *e.g. Nature* and *Science*, had developed news websites which report the latest advances of science at the level accessible to the general public. There are also plenty of science educational websites in English, which aggregate information from different sources like press-releases of universities and research institutes covering the scientific work of their employees.

In this regard Russia is lagging behind the West, especially after deterioration of educational traditions existed in the Soviet Union. Quite often a scientist learns about the achievements of his or her colleagues from a foreign journal because there are no press-releases issued by research institutes. It is even worse for people who are not involved in research — they have no idea about the activities inside their native research institutes.

Development of interactive network of widely-available educational resources in the Internet will allow popularization of biology, in particular among young people. Several popular science websites in Russian covering “the latest scientific advances” had appeared recently. Unfortunately, the majority of them contain only unprofessional translations from foreign resources of a similar kind, prepared by journalists lacking in scientific competence.

“Biomolecula.ru” ({{http://biomolecula.ru}}) is a young science educational website, which specifically covers the advances of contemporary physico-chemical biology (biophysics, biochemistry, molecular biology, bioinformatics, *etc.*). The main distinctive feature of the project is that it is run by people which are actively involved in scientific research — young postdocs and PhD students working in Russia and abroad. The most valuable content of “Biomolecula” is not short “news reports” on latest research but rather extensive and detailed reviews of various aspects of contemporary life science.

Online journal “Biomolecula” is a voluntary enterprise. The project exists and develops only because of the enthusiasm of its organizers and authors who have decided to spend their time on sharing their professional knowledge with the general reader. “Biomolecula” gladly welcomes new readers and authors — any scientist (a postdoctoral fellow, a PhD student, and even an undergraduate) who is eager to try his or her talent of science popularization could give it a try and join the community of “Biomolecula”.

CHROMOSOMICS, FROM NEW METHODS IN IMAGE ANALYSIS TO A NEW CHROMOSOME THEORY

Claussen U.
Institute of Human Genetics and Anthropology, Friedrich-Schiller-University,
Kollegiengasse 10, D-07740 Jena, Germany

Little is known about human interphase chromosomes mainly due to technical problems. Here, we analysed both the structure of chromosomes in the metaphase stage and in interphase nuclei using conventional GTG banding and high-resolution multicolour banding (MCB), which paints the total shape of chromosomes and creates a DNA-mediated, chromosome region-specific, pseudo-colored banding pattern at high resolution. The results show that metaphase chromosomes are highly elastic and can be stretched which leads to a splitting of GTG bands into their subbands. This phenomenon is mainly mediated by the chromosome harvesting procedure in which acetic acid and water (humidity) play a key role. The DNA-mediated shape, banding pattern and length of interphase chromosomes is nearly identical to that of metaphase chromosomes at all stages of the cell cycle. Furthermore, antibodies against histone modifications can be used to create a variety of different banding pattern on metaphase chromosomes. Based on the results obtained a new chromosome theory has been developed, **"The histone modification mediated hydration code of chromosomes"**.

MONTE CARLO SIMULATIONS OF THE 3D STRUCTURE OF THE PROTEIN BAK ASSOCIATED WITH MITOCHONDRIAL OUTER MEMBRANE

Davidovskii A.I., Veresov V.G.*
Institute of Biophysics and Cell Engineering NASB, Minsk, Belarus
e-mail: veres@biobel.bas-net.by
*Corresponding author

Motivation and Aim: Bax/Bak-mediated mitochondrial outer-membrane (MOM) permeabilization drives cell death (apoptosis) during development and tissue homeostasis [1]. The molecular mechanisms of apoptosis remain elusive largely due to the absence of high-resolution information about Bax/Bak 3D structures in the MOM-bound state. The aim of this study was to obtain the Bak three-dimensional structure in the model membrane mimicking the mitochondrial outer membrane.

Methods and Algorithms: The implicit membrane model for the MOM was used [2]. The Gouy-Chapman-Stern potential for the electrical double layer was applied outside the membrane slab. The ECEPP3 force field was used for calculations of intraprotein interactions. The free energies of partitioning the side chains of amino acids from water into the membrane was taken into account by the Kessel - Ben-Tal hydrophobicity scale combined with the Wimley-White interface hydrophobicity scale. The degree of solvation was taken into account by the use of the accessible surface areas with the use of the program GETAREA. The protein conformation was being described by the dihedral angles of the protein internal rotation, while Euler angles together with the coordinates of C_α atom of the first N-terminus peptide unit were used to describe the position of the protein as a whole relative to the membrane. The constraints obtained from functional and low-resolution structural data were applied to reduce the protein conformational space to be searched. A hierarchical simulation protocol combining coarse-grained discretization of conformational space with subsequent refinements [2] was applied to generate the protein conformation and its location in the membrane

Results: The dihedral angles and Cartesian coordinates of the atoms of the lowest-energy structure were obtained. The results show that in the vicinity of the membrane the release of the helix α9 from the hydrophobic pocket on the protein surface takes place followed by the partial unfolding of the protein and spontaneous insertion of the helices α5, α6 and α9 of Bak into the membrane.

Conclusion: The results obtained suggest that the strong association of Bak with the MOM, as compared to Bax, is caused by the absence of a significant interaction of the helix α9 with the remainder of the protein due, first, to a specific distribution of charges within Bak, and, second, to a shallow character, in the case of Bak, of the hydrophobic cleft on the protein surface.

Availability: The executable file and the PDB-file with atom coordinates of Bak integrated with the membrane are available from the authors upon the request.

References:

1. G. Kroemer et al. (2007) Mitochondrial membrane permeabilization in cell death, *Physiol. Rev.* **87:** 99-163.
2. V. G. Veresov, A. I. Davidovskii (2007) Monte Carlo simulations of tBid association with the mitochondrial outer membrane, *Eur. Biophys. J.* **37:** 19-33

PROTEIN CONNECTIVITY IN MOLECULAR-GENETIC NETWORKS DEPENDS ON PROTEIN SUSCEPTIBILITY TO SINGLE MUTATIONS

Demenkov P.S., Yarkova E.E., Ivanisenko T.V., Ivanisenko V.A. *
Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
Novosibirsk State University, Novosibirsk, Russia
Sobolev Institute of Mathematics, SB RAS, Novosibirsk, Russia
e-mail: salix@bionet.nsc.ru
* Corresponding author

Motivation and Aim: The characterization of protein interactions is essential for understanding biological systems. Within interaction networks, most proteins interact with few partners, while a small proportion of proteins, called hubs, participate in a large number of interactions and play a central role in biological processes. Thermodynamic stability is an important characteristic of proteins. Changes in thermodynamic stability upon mutations cause many diseases. Mutations can lead to increase or decrease in thermodynamic stability. Our aim was to identify relationships between human protein susceptibility to single mutations and connectivity of human proteins in a molecular-genetic interaction network.

Methods and Algorithms: For every human protein, we calculated the number of single mutations increasing, decreasing or not affecting thermodynamic stability. Protein susceptibility was defined as single mutations percent that decreased protein thermodynamic stability. The ProtStability program was used to predict changes in protein thermodynamic stability on primary structure. The ProtStability program is based on the «Modified KRAB» method. The ANDCell system was used to reconstruct associative network. Associative network consists of interactions and associations between human genes, proteins, microRNAs, metabolites, molecular processes, and pathways, cellular components.

Results: Susceptibility to single mutations within the 0.5-0.85 range was observed. Uniform distribution of protein susceptibility to single mutations was found for proteins with connectivity within the 1-200 range. The value of protein susceptibility to single mutations is tended to 74% for hubs. Protein susceptibility depends on the number of connections of protein in the associative network. Marginally essential proteins (hubs) have average susceptibility to single mutations.

Conclusion: Variation in susceptibility to single mutations decreased with the increase in the number of connections of protein in the network. Thus the susceptibility to single mutations for hubs varies in a close range. It may be assumed that for hubs exists the optimal value of susceptibility to single mutations that is necessary for maintenance of the flexible protein structure and noise immunity to mutations.

Work was supported in part by RFBR: 08-04-91313-IND_a and Government contract FASI №02.514.11.4065, interdisciplinary integrative project for basic research of the SB RAS № 115 and RAS presidium program “Molecular and cellular biology”, the grant “Systems biology: computer and experimental approaches.

A METHOD FOR THE ESTIMATION OF THE PARAMETERS OF THE LINEAR MODEL OF GENE NETWORK DYNAMICS

Demidenko V.G.*[1], *Podkolodnyy N.L.*[2, 3, 4]
[1] Institute of Computational Technologies, SB RAS, 630090, Russia;
[2] Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia;
[3] Institute of Computational Mathematics and Mathematical Geophysics, SB RAS, 630090, Russia;
[4] Novosibirsk State University, Novosibirsk, 630090, Russia
e-mail: pnl@bionet.nsc.ru
* Corresponding author

Motivation and Aim: The solution of the problem of the estimation of the parameters of the model for the gene network dynamics on the basis of microarray data is difficult and extremely timely [1]. The expression levels of a set of genes at a given time will influence the production of a given gene at a later time. The linear model is frequently used as a first approximation to complex nonlinear gene-gene interactions. The aim of this work was the development of a method of estimation of the parameters of a linear model of gene network dynamics on the basis of microarray data.

Methods and Algorithms: To describe gene-gene interactions, we used the following system of linear difference equations:

$$G_i(k+1) = \sum_{j=1}^{N} w_{ij} G_j(k) + w_{0i}\ ,\ i = 1,2,\ldots,N\ ,\ k = 1,2,\ldots,M-1\ ,$$

where $G_i(k)$ is the expression level of the i-th gene at the time point k; w_{ij} is the weight of the relation of the j-th gene to the i-th gene; w_{0i} is the free coefficient characterizing the external effect. The task is to define the weight coefficients w_{ij} and the free coefficients w_{0i} on the basis of the data on the expression level of a considered gene set.

This problem belongs to the inverse, which are traditionally difficult to solve because they are reduced to minimization of functionals of complex form. The method of problem solution we implemented is an extension of the idea suggested in [2] to solve the inverse problem for homogenous difference equations.

Results: We have developed a method and implemented its computer program for numerical solution of the inverse problem for a class of gene-gene interaction models based on microarray data, convergence and robustness of the method was demonstrated. The algorithm was also numerically analyzed. The main features of the proposed method are robustness and efficiency. At present, the method is modified for the case of the stationary linear model with variable external effect, the non stationary linear model, also for solution of the inverse problem high dimensionality.

References:

1. D. Stekel Microarray bioinformatics //Cambridge University Press, USA, 2003. – 263c.
2. M. Aoki Introduction to Optimization Techniques // M: Nauka, 1977.-344p. (Russian)
3. A.O. Egorshin On one method of evaluation of coefficients of modeling equations for sequences. // Sib. Zh. Ind. Mat., 2000, 3:2, - pp. 78-96. (Russian)

EXPERIMENTAL AND COMPUTER RESEARCH OF PROTEIN PATTERNS IN HEALTHY INDIVIDUALS AND PATIENTS WITH OVARIAN CANCER

Demidov E.A.[2]*, Govorun V.M.[1], Demina I.A.[1], Serebryakova M.V.[1], Yarkova E.E.[2], Ivanisenko V.A.[2]

[1] Centre "Bioengineering" RAS, Moscow, Russia

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

e-mail: scratch@gorodok.net

[*] Corresponding author

Motivation and Aim: Among all women oncological diseases ovarian cancer occurs in nearly 6% of all cases. Nevertheless this pathology is the main cause of lethal outcomes. The main aim of our work was to establish differences between protein patterns in healthy individuals and patients with ovarian cancer, to reveal proteins differentially represented in this two groups and to analyze their interactions.

Methods and Algorithms: To obtain minor compounds of serum proteins the IMAC method (immobilize metal affinity chromatography) was used. We applied MALDI to build the profile mass-specters of serum proteins. In these specters proteins are represented as peaks. Taking into account differences between protein profile mass-specters of healthy individuals and ovarian cancer patients the classification model for identification of normal and diseased state was built. This model is based on genetic algorithm.

To reconstruct the network of interactions between proteins ANDCell system was used. This computer system is designed for reconstructing network of interactions between different biological objects based on information automatically extracted from publication texts and factographic databases.

Results: The Mascot program revealed three proteins that were differentially represented in serum of normal individuals and ovarian cancer patients. They are: 1. human complement component C3, chain B; 2. kininogen-1 precursor; 3. complement component C4B.

These proteins together with those, mentioned in publications as ovarian cancer associated, were used for associative network reconstruction. This network allows us to suggest the possible mechanism of ovarian cancer pathogenesis.

Conclusion: We have built the classification model that is able to differentiate serum of normal individuals from this of ovarian cancer patients. The result of current study can be used for planning of further experiments and building of ovarian cancer fast diagnostic methods.

RASDB – REGULATION OF ALTERNATIVE SPLICING DATABASE

Denisov S.[1]*, Nurtdinov R.[1], Mazin P.[1], Kazakov A.[2], Kovaleva G.[2], Gelfand M.[2]
[1] Faculty of Bioengineering and Bioinformatics, M.V. Lomonosov Moscow State University, Moscow, Russia
[2] Institute for Information Transmission Problems RAS, Moscow, Russia
e-mail: stepan@bioinf.fbb.msu.ru
* Corresponding author

Motivation and Aim: The regulation of alternative splicing (AS) is an actual problem of modern molecular biology. Many cases of regulated splicing events are known, but the global picture is still obscure. The aim of our work is to collect all known cases of regulation of alternative splicing from published papers and to organize these data in formal way by creating a database.

Methods and Algorithms: From each paper we extracted the following information: all observed genes and mRNA isoforms (partial or full-length), cis-elements and trans-factors, which regulate expression of these isoforms. At the second step, we aligned genes and isoforms with the corresponding genome (we considered the following genomes: *H. sapiens*, *M. musculus*, *R. norvegicus*, *G. gallus* and *D. melanogaster*). This allowed us to merge all information about one gene from different papers. At the third stage we have developed software tools for manual curation of these data. Now we are preparing manual curation itself. We plan to create a web-interface to make the RASDB data publicly available.

Structure of the database: RASDB has the following structure. There are five main entry types: article, gene, isoform, isoform alignment, regulator (i.e. cis-element or trans-factor) and links between entries. The description of an isoform includes not only its exon-intron structure, but also the data about the expression of this isoform on different developmental stages and in different tissues and organs (if it has been presented in corresponding article). The names of organs, tissues and developmental stages for human, mouse and rat were organized in a tree using eVOC Ontologies ([1], {{http://www.evocontology.org/}}). We plan to perform analogous work for *Drosophila*. We also store the type of an experiment from a curated vocabulary.

There are two types of regulators: cis-elements and trans-factors. Trans-factors are molecules (proteins in most cases) which bind to some regions of a transcript (cis-elements), and then promote or block inclusion of , splicing out of introns, and the use of alternative sites . In our database, cis-elements have the following properties: DNA position, type of regulated event (cassette exon, alternative donor or acceptor site, retained intron etc.), the regulated isoform and the type of experiment. Trans-factors are described by the molecule type (protein or RNA) and GenBank Accession.

Currently RASDB contains 633 cis-elements, 469 trans-factors, which regulates splicing of 459 genes. The RASDB database can be used for the analysis of regulatory motifs and trans-factors or for understanding the trends and characteristic properties of the evolution of regulated AS events.

We are grateful to members of our scientific group and external annotators for data input.

This work is suppported by grants from RFBR (07-04-00343), HHMI (55005610) and program "Molecular and Cellular Biology" RAS

References:

1. J.Kelso et al. (2003) eVOC: a controlled vocabulary for unifying gene expression data, *Genome Res.,* **13(6A):**1222–1230.

COMPUTER SIMULATION OF C.ELEGANS MUSCULAR SYSTEM AND NEURAL NETWORK

Dibert A.A., Palyanov A.Yu. *
Novosibirsk State University, Novosibirsk, Russia.
e-mail: shamdor@mail.ru**,** palyanov.andrey@gmail.com
*Corresponding author

Motivation and Aim: Investigation of nervous system structure and functioning is one of the most interesting and complicated problems. Functional computer model of a nervous system, which reproduces the properties of the original with high accuracy, would be an evidence of high understanding level of processes taking place in it. Reproducing the architecture of some real neural network seems to be a good approach to start with. The mammal brain and even brains of simpler organisms is too complex both to determine the positions of all the neurons and connections between them and to simulate it on modern computers. Moreover, although a lot of different neuron models have been proposed, it is difficult to estimate how close to reality they are.

C.Elegans, free-living soil nematode, is one of the model organisms, widely used and extensively studied by biologists. It is the only organism for which neural network architecture – positions of its neurons and connections between them - is almost completely known. Its nervous system consists of 302 neurons, over 5000 synapses, more than 2000 neuromuscular junctions and these elements are invariant for individuals of the same sex.

Taking into consideration aforesaid, simulation of C.Elegans nervous system seem to be one of the most actual and necessary task. Small size of neural network will allow us to make calculations in reasonable time using modern computers. Besides the model of the nervous system model, it is very important to develop the model of organism's body including muscles and receptors in three dimensional physical environment, which will provide sensory input and feedback to the working nervous system and allow to observe organism's behavior.

Methods and algorithms: The program is written in C++ programming language using free graphics library OpenGL for visualization. Simple model of input signal summation with adjustable actuation threshold is used for realization of interaction between neurons. Neurons can be both activating and inhibitory. Information on neuron thresholds and interneuron connection weights is unknown, so we plan to use both experimental research data and algorithms of neural networks learning to approximate adequate values.

Results: At the present moment 3D-simulator of C.Elegans muscular system has been realized, which includes, as a real muscular system, 4 longitudinal groups of muscles, which can be affected by signals from motoneurons of the nervous system. For calculations of dynamical physical model we take into account the supporting force, the friction force, the muscle tension, gravity and the resistance of medium. Also the C++ classes for loading and simulating given neural network are realized. We use the experimental data, describing C.Elegans nervous system, which are available at www.wormatlas.org/MoW_built0.92/toc.html (The Mind of a Worm Project) which are loaded into our program.

Conclusion: The work which is already done is the first stage of the project. Further we plan to elaborate the more realistic physical model, to tune the most investigated fragments of the nervous system, including motory circuit, which is responsible for the movements of the organism, to add simulation of tactile receptors to the model and finally to adjust weights and thresholds for all neurons and connections between them in such a way, which will result in adequate behavior of the animal.

UNIVERSAL METHOD FOR REVEALING OF DRUG-RESISTANT FORMS OF HIV-1

Dmitrienko E.V.[1, 2], Pyshnaya I.A.[1], Repkova M.N.[1], Levina A.S.[1], Gashnikova Y.S.[3], Kabilov M.R.[1], Pyshnyi D.V.[1, 2*], Zarytova V.F.[1]

[1] Institute of Chemical biology and Fundamental medicine, SB RAS, Lavrentyev pr. 8, Novosibirsk 630090, Russia

[2] Novosibirsk State University, Russia

[3] State куыуфкср center of virology and biotechnology "Vector", Koltsovo, Russia

*e-mail: pyshnyi@niboch.nsc.ru

Motivation and Aim: The line of the present investigation implies the solution of the fundamental scientific problem, viz., the development of key approaches providing high selectivity of the interaction between nucleic acids as well as high accuracy of revealing and affecting a certain genetic material. The goal of this project is the elaboration of the unique technology of revealing point mutations in the presence of single nucleotide polymorph substitutions in the HIV-1 genome, which determine resistance of the virus to antiviral drugs of the nucleoside group.

Methods and Algorithms: The method is based on the registration of the point mutations responsible for the drug resistance in the presence of insignificant polymorphic substitutions in DNA localized in the oligonucleotide probe binding site. Unique miniprobes, namely, tandems of short oligonucleotides (composite tandems or cross-linked by a non-nucleotide spacer tandems) were used as oligonucleotide probes.

Results: A new strategy of revealing drag resistant forms of HIV-1 has been proposed. Degeneracy of DNA was taken into account while synthesizing oligonucleotide probes. The revealing of point mutation in the analyzed DNA was performed by means of ligation of the miniprobes. Analysis was carried out using RT PCR fragments of gene *pol* HIV-1 isolated the blood of HIV-infected patients. Among insignificant substitutions, these fragments contained those in codons 62, 67, 70, 184, and 215 causing the resistance of viruses to antiretroviral therapy. The use of the proposed miniprobes was shown to enhance selectivity of revealing important point mutations in DNA even in the presence of neighboring SNP localized in the probe binding site. The obtained results are the basis for the development of the test-systems for revealing mutant HIV-1 strains resistant to nucleoside antiviral drugs. Such test systems are presented by the set of relatively inexpensive DNA chips and miniprobes, provided the biotin-streptavidin system and chromogenic substrates are used for the colorimetric detection of the signal ligation products.

Conclusion and Availability: The proposed method has no analogs both in our country and abroad, exceeds the known approaches in terms of its sensitivity to one-nucleotide substitutions, and allows diagnostics to be carried out during several hours after the probe preparation. The developed strategy of revealing point mutations in the presence of polymorphism in the DNA structure can be utilized for the detection of other drug-resistant viruses and bacteria.

This work was supported by MCB Program of RAS (10.6), integration grant of SB RAS (55, 73), RFBR grant (06-04-49263).

GENTOO PENGUIN'S POLYMORPHISM ON MOLECULAR-GENETIC LEVEL

Dranitsina A.S.[1]*, Telegeev G.D.[2], Bezrukov V.F.[3]
[1] Scientific Research Institute of Physiology Named After Academician Peter Bogach of the Kyiv Taras Shevchenko University, Kyiv, Ukraine
[2] Institute of Molecular Biology and Genetics of NAS of Ukraine, Kyiv, Ukraine
[3] National Taras Shevchenko University of Kyiv, Kyiv, Ukraine
* Corresponding author

Motivation and Aim: Gentoo (*Pygoscelis papua*) is one of the species-indicator for the evaluation of the impact of global environmental changes in the Antarctic ecosystem. Genetic diversity of Gentoo penguin populations, telomere length and sex as bioindicator in investigation of Gentoo populations in relation to Antarctica environmental state have been analyzed in the present paper.

Methods and Algorithms: RAPD-PCR was carried out as the modified method of Operon Technologies (Alameda, CA, USA) with series of random primers 10 bp each. Sex identification and microsatellite loci analyses have been carried out by PCR with specific primers. The special telomeric probe was constructed by ligating telomeric repeats with the following cloning in *pUC19* and was labeled by P^{32}dCTP or dUTP digoxigenin. Data analysis: RFLP Scan 3.12. (Scanalytics), POPGENE version 1.32, PHYLIP package, version 3.63, GraphPad Prism 4.03 (GraphPad Software Inc., USA).

Results: RAPD-analysis revealed the levels of polymorphism in Gentoo at Petermann Island (from 23,5 to 42,9%) and from Livingston Island (from 52,9 to 57,1%). The high level of relationship between two Gentoo populations with a lack of significant genetic differentiation between them (Fst=0,069) was demonstrated, despite substantial levels of genetic variation. So it was confirmed that these two populations belong to the same subspecies (*Pygoscelis papua ellsworthi*).

Female/male indexes (proportion of female) were: 0,336 for population of Petermann Island and 0,398 for Livingston Island. Only 20 penguins from Petermann were females among 98 chicks ($\chi 2=4,1$, df=1, p=0,04), so Fisher's hypothesis was not satisfied (we found significant deviation from predicted binomial distribution) and perhaps the determined sex ratio was under Trivers and Willard's and Charnov's models.

AM12 and RM6 microsatellite loci in populations of Gentoo penguins were absolutely monomorphic with only one allele in both populations. RM3 demonstrated two alleles presented in this locus – 221 b.p. and the new allele - 217 b.p., which has not been described for RM3 to date. We also could confirm that these two populations belong to the same subspecies (*Pygoscelis papua ellsworthi*), Fst<0,04.

The average observed telomere length of Gentoo penguins was 5950±1537 b.p. for adult specimens and 8100±949,1 b.p. for chicks (p<0,0001). The maximum telomere length for adult specimens was 8000 b.p., the minimum – 3100 b.p. The maximum telomere length of Gentoo chicks was approximately 9000 b.p., the minimum – 6200 b.p. Thus we can draw a conclusion that the telomere length of 9000 b.p. is the maximum for the specimens of this species.

Conclusion: Both telomere length ("population age") and sex could be used as appropriate bioindicators in research of Gentoo populations in relation to Antarctica environmental state in monitoring programs of Anarctica ecosystems.

APPLICATION OF NONMETRIC MULTIDIMENSIONAL SCALING FOR ANALYSIS OF CROSS-PLATFORM GENE EXPRESSION MICROARRAY DATA

Efimov V.M. *, Katokhin A.V.
Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
e-mail: vmefimov@bionet.nsc.ru
* Corresponding author

Motivation and Aim: Incompatibility of results obtained upon analysis of signal intensities for same probes presented on different platforms is main problems met during analysis of microarray data on gene expression. However it is possible to calculate a matrix of dissimilarities between samples across all probes set for homologous experiments with each platform. The matrices for different platforms should be similar (perhaps with accuracy to nonlinear monotonous transformation). Nonlinearity is eliminated by dissimilarities ranking. Hence an integrated dissimilarity matrix could be built and gene expression profiles (GEPs) produced by each platform and corresponding to specific directions in sample space could be choosen.

Methods and Algorithms: We analysed two sets of GEP data for 31 peripheral blood samples from patients with Huntington's disease, presymptomatic individuals, and normal cases (1). Upon filtration procedure CodeLink set contained 17526 complete GEPs and Affymetrix set - 22283 ones. The sets were logarithmic transformated, centered and normalized. For each sets an Euclidean distance matrix between the samples was calculated with replacement of distances by their ranks. The matrices were averaged and processed by nonmetric multidimensional scaling (2). So every starting set element became assigned to a point in Euclidean space of low dimension (e.g. plane). Then the points shifted such that the matrix of distances be fit well to the matrix of dissimilarities between elements with using a coefficient of rank correlation as fitness criterium.

Results: All samples partitioned to four classes, corresponding to three mentioned groups and intermediated one between presymptomatic and healthy individuals (subnormal, assigned as healthy by (1)). The four partitions occupied distinct regions on the plane. Noteworthy, the information on initial affiliation of individuals with each group was not used, the samples were unsupervisely partitioned according GEPs.

Conclusion: Analysis of position of the groups allowed to suppose a possibility of consecutive transition along circular path: healthy => subnormal => presymptomatic => patients, with specific GEPs corresponding to each group. To reveal them the correlation coefficients were calculated between the GEPs and directions in plane from centroid of all points set towards centroids of each group. The GEPs with highest possible correlations by module for each group were selected. Candidate genes (1) suitable for Huntington's disease diagnosis were present in patients group gene lists generated for both sets.

References:

1. F. Borovecki *et al.* (2005) Genome-wide expression profiling of human blood reveals biomarkers for Huntington's disease. *Proc Natl Acad Sci U S A.* **102:** 11023-11028.
2. Y.H. Taguchi, Y. Oono (2004) Relational patterns of gene expression via non-metric multidimensional scaling analysis. *Bioinformatics.* **21:** 730-740.

FEATURES OF THE MICROSATELLITE LOCI IN POPULATIONS OF SOME KINDS OF ADDERS (VIPERIDAE, VIPERA)

Efimov R.V.[1]*, Zavialov E.V.[1], Tabachishan V.G.[2]
[1] Saratov State University, Saratov, Russia
[2] Saratov branch of A.N. Severtsov Institute of Ecology and Evolution RAS, Saratov, Russia
e-mail: EfimovRV@Rambler.ru
* Corresponding author

Motivation and Aim: Systematic position of *Vipera nikolskii* (Vedmederja, Grubant et Rudaeva, 1986), living in the forest-steppe and northern part of the steppe zone of the Eastern Europe, is insufficiently studied. Use of the morphological methods of the analysis does not allow to solve the given problem [1]. In this connection for studying the taxonomic status of *Vipera nikolskii* it is necessary to involve the molecular genetic techniques based on the analysis of the microsatellite sequences.

Methods and Algorithms: In order to study the systematic status of *Vipera nikolskii* we have chosen the microsatellite loci 7-87, 5 and 7, used for revealing the genetic relationship of *Crotalus horridus, Nerodia fasciata fasciata* [2,3].

Results: The total 61 samples of *Vipera nikolskii* and *Vipera berus* from various habitat of the Saratov, Samara, Penza, Tula, Novgorod and Perm areas, the Chuvash Republic and the Republic of Mordovia were selected for research.

We found 10 alleles for the microsatellite loci 7-87, which have size from 152 to 192 bp. The level of the heterozygosity on this microsatellite loci was 72 %. It should be noted that the allele in the size of 152 bp was present only in the populations of *Vipera nikolskii* from the Saratov region, and the allele in the size of 192 bp was present only in the population of *Vipera berus* from the Perm area.

The *Vipera berus* from the Tula and the Novgorod regions were characterized by only two alleles 176 and 184 bp. All other populations had a full set of microsatellites that testifies to existence in the given territory of the transitive forms. Loci 5 and 7 were completely monomorphic of all adder used in the analysis.

Conclusion and Availability: This work is the first attempt to analyze the microsatellite sequence of *Vipera nikolskii*. In this connection the further comparative analysis of the microsatellite sequence of investigated species of adder from various regions of the area is represented perspective.

References:

1. W.G. Tabatschischin et al. (2004) Zur präzisierung der südlichen Grenze des Verbrei-tungsareals der Waldsteppenotter (*Vipera nikolskii*) im europäischen Teil Russlands, Mauritiana, 19: 83-85.
2. X. Villarreal, et al. (1996) Isolation and Characterization of Microsatellite Loci for Use in Population Genetic Analysis in the Timber Rattlesnake, *Crotalus horridus,* The Journal of Heredity, 87: 152-155.
3. R. Melanie et al. (1999) Microgeographic population genetic structure in the northern water snake, *Nerodia sipedon sipedon* detected using microsatellite DNA loci, Molecular ecology, 8: 329-333.

MODERN COMPUTATIONAL PHARMACOLOGY: MOVING TOWARD BIOMEMBRANES

Efremov R.G.*, Nolde D.E., Polyansky A.A., Novoseletsky V.N., Volynsky P.E.
Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Russian Academy of Sciences, ul. Miklukho-Maklaya 16/10, GSP Moscow, 117997 Russia.
* Corresponding author. e-mail: efremov@nmr.ru

Motivation and aim: Nowadays, computational (or *in silico*) technologies are widely used in design of new biologically active compounds. Along with "traditional" objects of a pharmacological interest (like water-soluble proteins, nucleic acids, and their complexes), a growing attention is now attached to targets in cell membranes. These are the membrane-bound and membrane-active peptides and proteins (MPs) as well as the lipid bilayer itself. MPs play a crucial role in numerous cell processes, such as signaling, ion conductance, fusion, and others. Many of them act as highly specific and efficient drugs and drug targets. Because of experimental difficulties with structural characterization of MPs, computer simulations became an important alternative source of biologically relevant information for such supramolecular systems.

Methods and Algorithms: Structural/dynamic properties of MPs with diverse fold (α-helical, β-structural), mode of membrane binding, and biological activities were assessed *via* simulations with implicit and/or explicit theoretical models of membranes. Among the objects under study were antimicrobial and fusogenic peptides, cardiotoxins, GPCRs, artificial peptides targeting transmembrane helices of MPs, *etc*. [1-3]. A new computational approach was proposed to study behavior of MPs in different membrane-mimic media. The approach combines in a self-consistent manner Monte Carlo conformational search in implicit hydrophobic slabs, molecular dynamics in hydrated full-atom lipid bilayers and micelles, molecular hydrophobicity potential analysis, homology modeling, *etc*.

Results: The predictive power of the computational protocols was proven *via* testing the modeling results against high-resolution experimental data obtained for several antimicrobial and fusogenic peptides, transmembrane helix-helix dimers, and so on. Furthermore, based on the theoretical data, a number of MPps with "improved" biological activities were elaborated.

Conclusion: In silico technologies represent an integral part of modern computer-aided design of novel biologically active molecules – potential drugs - acting on targets in cell membranes.

References:

1. R.G. Efremov *et al*. (2004) Peptides and proteins in membranes: what can we learn *via* computer simulations? *Curr. Med. Chem.,* **11**: 2421-2442.
2. R.G. Efremov *et al.* (2007). Molecular lipophilicity in protein modeling and drug design, *Curr. Med. Chem.*, **14**: 393-415.
3. Ya.A. Vereshaga *et al.* (2007). Specificity of helix packing in transmembrane dimer of the cell death factor BNIP3: a molecular modeling study. *Proteins*, **69**: 309-325.

THE NEW ALGORYTHM FOR PHYLOGENETIC RECONSTRUCTION OF NON-RECOMBINING DNA SEQUENCES

Eltsov N.P.*, Volodko N.V.
Laboratory of Human Molecular Genetics, Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
e-mail: eltsovnp@bionet.nsc.ru

Motivation and Goals: The mitochondrial genome has been the most widely used system for the investigation of the evolutionary history of our species. It has become a system of choice because of its high rate of sequence divergence and because of its uniparental, maternal inheritance. With the advent of human population genomics [1] and rapid accumulation of complete mtDNA sequences, it has become increasingly important to quickly and comprehensively analyze the data available.

Methods and Algorithms: We propose a novel maximum parsimony-based algorithm for reconstruction of phylogeny of non-recombining DNA sequences. This algorithm includes three consecutive steps. 1) sorting and identification of recurrent mutations (the ones that do not allow for unambiguous phylogeny reconstruction); 2) analysis of recurrent mutations and identification of the most plausible parallel mutations; 3) parallelization of these mutations. When aligning the DNA sequences, we also applied a novel optimized algorithm for weighted alignment.

Results: The algorithm designed was used in mtPhyl. This software package allows analyzing rapidly human entire mtDNA sequences. The mtPhyl identifies the mutated region, aminoacid replacements, and calculates the coalescence time for the most recent common ancestor. In addition, it sorts out sequences in accord with parameters as outlined by user, and enables us to estimate the natural selection. The output can be easily converted into any formats used by popular programs such as Arleqin, DnaSP, etc. The mtPhyl appears to be a unique package which can be used as a standalone tool and as an accessory program for preliminary data analysis.

Conclusion: mtPhyl based on a new algorithm to reconstruct molecular phylogeny represents a timely advance, since the advent of cheaper sequencing methods has generated an excess of sequence data, and there is an urgent need to perform their automatic analysis.

Availability: Demo version of mtPhyl is available from the authors upon request and at http://www.bionet.nsc.ru/labs/mtgenome/programs.html.

References:

1. S.B. Hedges (2000) Human evolution. A start for population genomics, *Nature*, **408**: 652-653.

THE PARADIGM OF KNOWLEDGE DISCOVERY IN LIFE SCIENCES

Famili A.
Institute for Information Technology
National Research Council of Canada
Ottawa

Abstract: The need for knowledge discovery in life sciences has resulted from several trends among which three are the most important ones. These are: (i) methods, tools and technologies that allow one to get a holistic understanding of the entire genome, (ii) the recent reality that proper understanding of biological processes needs to be an integration of several biological research directions, such as genomics, metablomics, proteomics, etc, and (iii) the basic reality that knowledge discovery would help us in many directions such as functional omics, better prognostics and diagnostics in healthcare, intelligent drug discovery, personalized medicine and many more. Essentially, knowledge discovery is the process of developing and applying strategies to discover useful and ideally all previously unknown knowledge from historical or real-time data. Applied to the life sciences domain, knowledge discovery processes will help in various research and development activities, such as (i) studying data quality for possible anomalous or questionable biological data/experiments, (ii) identifying relationships between genes/proteins and their functions based on time-series or other high throughput biological profiles, (iii) investigating genes/proteins responses to treatments/environment under various conditions such as *in-vitro* or *in-vivo* studies, and (iv) discovering predictive/descriptive models for accurate diagnosis/classifications based on biological expression profiles among two or more classes.

This talk consists of three parts. In part one, we provide an overview of knowledge discovery focusing on bioinformatics domain and describe the main motivations for developing and applying knowledge discovery methods to analyze life sciences data. In part two of this talk, we briefly describe a few of our case studies where we have analyzed high throughput biological data. These are cases in which real biological data sets (obtained from public or private sources) have been used for tasks such as gene function identification and gene response analysis. In the last part of this talk, we will describe the overall trends in the paradigm of knowledge discovery in life sciences.

Short Bio: Dr. A. Famili is a Senior Research Scientist, Group Leader for Knowledge Discovery Group and a leading data mining expert working at the Institute for Information Technology (IIT) of the National Research Council of Canada, where he has been for the last 23 years. Prior to joining NRC, he worked in industry for 3 years. Dr. Famili has been actively involved in the field of Artificial Intelligence, Data Mining and BioInformatics and successful application of these technologies. He has a strong data mining and bioinformatics team within IIT that is currently engaged in unique research and development in data mining for genomics, proteomics and health care. His research has been on data mining, machine learning and bioinformatics and their applications to real world problems in various data rich environments, such as life sciences. Dr. Famili has edited two books, has published over 45 articles in the area of data mining and AI and has a US data mining patent. He has organized many workshops and has been involved in a number of data mining and AI conferences and has extensive collaboration with NRC Institutes and a number of other research institutes in Canada and Europe. He is also on the editorial board of four scientific journals and an adjunct professor at SITE (School of Information Technology and Engineering), and The Institute of System Biology, at the University of Ottawa.

DIFFENTIAL DEMOGRAPHIC IMPACT OF GEOLOGICAL EVENTS ON MOLLUSCAN SPECIES OF DIFFERENT ECOLOGICAL AFFILITIES

Fazalova V.*, Ivanova Z., Sherbakov D.
Limnological Institute Siberian Branch of the Russian Academy of Sciences,
Irkutsk 664003, Russia
* Corresponding author

Motivation and Aim: Previous studies on genetic variation of invertebrates from the Lake Baikal revealed variation of population dynamics caused by geological events (tectonical shifts), changes in global climate and related shifts in segmentation. Maakia (Eubaicalia) herderiana is one of the molluscan species dominating the stone littoral of the lake as opposed to other Baikalian species inhabiting sand bottom. This species varies in shell sculpture: from ribbed to smooth shells. The aim of this study was to compare genetic and morphological differentiation of this molluscan species and analyze possible influence of past geological events on demographic of and genetic variation.

Methods and Algorithms: We used mitochondrial COI DNA sequences to infer the pattern of genetic diversity in M. herderiana inhabiting the southern shore of the lake Baikal. We have analyzed 68 samples from 7 localities. In some of them ribbed and smooth shell morphs were found in sympatry. To estimate population-genetic parameters such as population size and population growth rate we used program LAMARC version 2.1.2b [1].

We used individual-based models, developed in line with [2] in order to simulate patterns of sequence diversity resulting different demographic scenarios.

Results: The observed proportions of nucleotide substitutions among M. herderiana mtDNA COI sequences were 0.0 – 0.7%. We did not reveal any correlation between the shell sculpture and the slight genetic differentiation. This suggests that organisms from all sampled localities form a single population. Observed patterns of molecular and morphological diversity of M. herderiana have been compared to simulated patters and similar results obtained by others for ecologically dissimilar benthic animals inhabiting the same area of lake Baikal.

Conclusions: No animal populations of stable demography have been observed so far at South-Western shore slope of Lake Baikal. M. herderiana is a good example of a rock-dweller which is likely to use opportunities created by tectonic events know to occur during last 100 000 years in this area (Maz and Sherbakov; pers. comm.). Here both simulations and experimental approaches have been used to elucidate demographic history of M. herderiana and some benthic species and correlated to known tectonic catastrophes.

References:

1. Kuhner, M K. (2006) LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters, Bioinformatics, 22: 768–770.
2. Semovski S.V. et al (2004). Simulating the evolution of neutrally evolving sequences in a population under environmental changes, Ecological Modelling, 176: 99-107.

BIOLOGICAL SPECTRA ANALYSIS: LINKING BIOLOGICAL ACTIVITY TO ADMET PROFILES

Fedichev P.*, Vinnik A.
Quantum Pharmaceuticals, Kosmonavta Volkova street, 6A, Moscow, Russian Federation.
* Corresponding author, e-mail: peter.fedichev@q-pharm.com

Motivation and Aim: The inability to accurately predict toxicity early in drug development resulted in $8 billion losses in the pharmaceutical industry in 2003, approximately one-third of the all drug failures costs in one year . A recent example is Merck's withdrawal of the blockbuster drug Vioxx, which caused the company's stock to plunge 25% in one day . That is why computer aided toxicology prediction is of growing interest both to the industry and for government regulators, who have recently issued several reports calling for application of NEW approaches in assessing drug safety.

Methods and algorithms: Recently it was observed that experimental values of molecular activities against a large proteins set can be used for predicting broad biological effects . In this investigation we take advantage of this concept and develop a novel qualitative method for ADMET properties prediction. Using our in-house docking and binding free energies prediction software (QUANTUM), we calculate inhibition constants for about *1200* small molecules, mostly drugs, against a diversified *510* human proteins set. Using publicly available ADMET information for every molecule we establish relationships between the binding constants and various ADMET properties (endpoints).

Results: The analysis of binding affinities profile for each of the molecules led to establishment of a relation between various ADME and toxicological properties of a molecule and its activity against selected representatives of approximately *50* protein-families.

Conclusion: We have shown that a new approach, linking the computed Biological Spectra with ADMET properties of small molecules. This method is very flexible, can use both in-vitro and in silico assays, and has, in principle, a potential to reduce or substitute for animal testing. It is capable of incorporating diverse databases of molecules with different biological activities, and can have significant implications for biology and drug discovery.

Availability: The software demo, q-Tox, is available from Quantum Pharmaceuticals website, http://www.q-pharm.com

FACTORS INVOLVED IN REGULATION OF L1 RETROTRANSPOSONS EXPRESSION

Fedorov A.V.*, Podgornaya O.I.
Institute of Cytology RAS, Saint-Petersburg, Russia
e-mail: an_tn@mail.ru
* Corresponding author

Motivation and Aim: L1 elements belong to the superfamily of autonomous retrotransposons and constitutes roughly 10-20% of mammalian genome. L1 elements encode two proteins: one is RNA-binding protein that specifically interacts with L1 transcripts to form ribonucleoprotein particles and the second one possesses reverse transcriptase and endonuclease activities. The both proteins are assumed to be required for retrotransposition of LINEs. Transcription of full-length RNA, the first step in the L1 retrotransposition, provides a template for synthesis of both L1-encoded proteins and a DNA copy. Coexpression of full-length sense strand LINE1 RNA and LINE1 encoded protein can be detected only at definite stages of germ cell development, and in case of some cancers. Thus accurate regulation of L1 transcription plays important role in the retrotransposition control and genome fate; however mechanisms of such regulation remain largely unknown.

Methods and Algorithms: We investigated factors involved in regulation of L1 elements expression with the help of northern and southern blotting and computer analysis of L1 promoter regions structure using MAR-Wiz 1.0 program (http://www.futuresoft.org/MAR-Wiz/) and DNA Curvature Analysis Software (http://www.lfd.uci.edu/~gohlke/curve/).

Results: In our study we characterised pattern of expression of rat L1 elements, their promoter structure and methylation state and identified transcription factors specifically binding to this promoter *in vitro*.

Conclusion: We conclude that factors such as DNA bending and association with nuclear matrix seem not to be involved in regulation of rat L1 transcription. In contrast cell type specific methylation in cooperation with ubiquitous transcription factors could modulate transcriptional activity of rat L1 promoter. Heterodisperse sense and antisense L1 transcripts which were revealed in liver cells, most likely are not translated and unable to participate in the retrotransposition of L1. Finally, we detected small sense and antisense RNAs homologous to rat L1 promoter region in testis cells. Thus we suggest that complimentary L1 transcripts could form double stranded RNA, which are processed to small RNA and hypothetically participate in RNA-dependent silencing of L1 elements.

A PROTOCOL TO DEMONSTRATE SEQUENCE MATCHING

Fedyukovych V.E.*[1], Sharapov V.G.[2]
[1] GlobalLogic Ukraine, Kiev, Ukraine
[2] Physical-technical institute of NTUU "KPI", Kiev, Ukraine
[*] Corresponding author: e-mail: vf@unity.net

Motivation and Aim: A need to resolve conflicting interests related to private data availability is the reason for our research. We consider a problem of making verifiable statements about properties of sequences widely used in bioinformatics. Our major targets are nucleotide sequences and conclusions resulting from DNA analysis. We focus on protocols to produce verifiable statements involving approximate matching without the need to disclosure private information. We stress that there was a very limited development supporting verifiable statements related to approximate matching.

Methods and Algorithms: We introduce a new proof of knowledge protocol for K-way substring statement, a related verification method, and a polynomial representation for strings. We use a novel challenge-response system designed to prove properties of polynomials. We use Schnorr proof of knowledge protocol and Pedersen commitment scheme to bootstrap challenge-response system from public information.

A polynomial representation for multisets was introduced for a set reconciliation protocol,

and was later extended to represent directed graphs. Verification method using responses of Schnorr protocol was introduced for a set similarity protocol, and was also used for protocols to prove knowledge of a Goppa codeword, graph isomorphism a k-colourability. Challenge-response system from Schwartz-Zippel lemma was introduced for set similarity protocol, and could be considered an extension of Schnorr protocol. An efficient protocol to prove knowledge of a logarithm in a finite group was introduced by Schnorr for an electronic signature scheme. A commitment scheme was introduced by Pedersen to facilitate protocols to prove practical statements about committed data. Commitments and proof of knowledge protocols were used to design democratic voting and electronic cash systems, verify correctness of computations with shared data.

Results: Protocol developed allows a proving party to demonstrate validity of a statement regarding his data only using public information about his data. Our protocol was designed to demonstrate that at least K copies of a pattern string are embedded in a host string. Our protocol does not leak information about both strings and matching locations, other than string lengths and the statement shown. It was shown protocol developed is a special honest verifier zero knowledge argument of knowledge; probability estimate was given for a honest Verifier to accept for any Prover trying to show a false statement; this probability was shown to be exponentially small in group order bitsize.

Conclusion: We consider substring statement to be a convenient model useful to approach specific practical protocols. We expect new protocols to appear addressing specific statements and conclusions like identification, paternity test, disease susceptibility etc.

Availability: Basic protocol for set similarity was implemented to show protocols are practical and useable. Additional information supporting software development is available at http://vf.org.ua/string_matching/

MACHINE LEARNING TECHNIQUES FOR NUCLEAR HORMONE RESPONSE ELEMENT PREDICTION

Fourati A.*, Choura M., Aifa S., Rebai A.
Bioinformatics Unit, Centre of Biotechnology of Sfax, Sfax, Tunisia
e-mail: ahmed.fourati@cbs.rnrt.tn
* Corresponding author

Motivation and Aim: One of the challenges of bioinformatics in genome analysis is to predict particular motifs within DNA sequences serving as trans-regulation sites. Nuclear Hormone Response Elements (NHRE) are among the most important response elements adopted by Eukaryotic regulation of genetic expression.

Methods and Algorithms: An NHRE is a DNA sequence characterized by its palindromic structure with 3 specific features: the sequence of the base pairs in the half-site, the number of base pairs between the half-sites, the relative orientation of the two half-sites.

These sequences can be predicted through different methods such as Weight Matrix (WM), Markov Model (MM) and Bayesian Networks (BN). Although MM has been shown to be the best methods to predict DNA motifs in general, their application to NHRE prediction was not very successful due to the particular properties of the NHRE.

Results: In this work, we first outline the performance of the three methods and how they can be adapted to NHRE prediction. Then we present two programs that implement WM and BN methods. We used these programs to perform a genome wide search of NHRE in the human genome.

Conclusion: This allowed us to evaluate the performance of the two methods and to compare them to each other.

We finally propose a combination of BN and WM to improve the precision of the prediction.

Availability: NHRE finder will be soon available as academic tool on the CBS web site (www.cbs.rnrt.tn)

References:

1. W. S. Dalton and S. H. Friend "Cancer Biomarkers—An Invitation to the Table" Science 26 May 2006 Vol. 312. no. 5777, pp. 1165 - 1168
2. L. E. Tavera-Mendoza, S. Mader, and J. H. White "Genome-wide approaches for identification of nuclear receptor target genes(Published online 2006 July 7)" Nucl Recept Signal. 2006; 4: e018.
3. V. B. Bajic, S. L. Tan, A. Chong, S. Tang, A. Ström, J. Å. Gustafsson, C. Y. Lin and E. T. Liu "Dragon ERE Finder version 2: a tool for accurate detection and analysis of estrogen response elements in vertebrate genomes" Nucleic Acids Research, 2003, Vol. 31, No. 13 3605-3607
4. Sandelin, W. Wasserman "Prediction of nuclear hormone receptor response elements" Mol Endocrinol. 2005 Mar;19(3):595-606
5. M. J. Malti "PREDICTION OF RESPONSE ELEMENTS OF NUCLEAR HORMONE RECEPTORS" M.S. thesis South African National Bioinformatics Institute (SANBI) University of Western Cap June 20, 2006
6. V. Matys*, O. V. Kel-Margoulis, E. Fricke, I. Liebich, S. Land, A. Barre-Dirrie, I. Reuter, D. Chekmenev, M. Krull, K. Hornischer, N. Voss, P. Stegmaier, B. Lewicki-Potapov, H. Saxel, A. E. Kel and E. Wingender "TRANSFAC® and its module TRANSCompel®: transcriptional gene regulation in eukaryotes" Nucleic Acids Research, 2006, Vol. 34, Database issue D108-D110
7. J. Sladeczek, A. J. Hartemink, and J. Robinson "Bayesian Network Inference with Java Objects"Department of Computer Science Duke University http://www.cs.duke.edu/~amink/software/banjo/
8. G. Nuel. Pattern statistics on Markov chains and sensitivity to parameter estimation. Algorithms Mol Biol. 2006, 1:17.
9. Deyneko IV, Kel AE, Bloecker H, Kauer G. Signal-theoretical DNA similarity measure revealing unexpected similarities of E. coli promoters. In Silico Biol. 2005; 5: 547-555.

HUMAN AND MOUSE GENOMES, $(A)_nB$-REPEATS AND *ALU*-LIKE ELEMENTS

Fridman M.V.[1]*, Oparina N.J.[2], Makeev V.J.[1,2]
[1]Institute of genetics and selection of industrial microorganisms, GosNIIgenetika, Moscow, Russia
[2]Engelhardt Institute of Molecular Biology, RAS, Moscow, Russia
e-mail: marina-free@mail.ru
[*]Corresponding author

Motivation and Aim Microsatellites are one of the most dynamic class of genome sequences responsible for many functionally important types of genome variability [1,2]. Our objective was using full-genome data to find dominating types of microsatellites with periods 4-6 in human and mouse genomes and determine factors explaining their abundance.

Methods and Algorithms We used TRF-based "Simple repeat" datatrack associated with UCSC GenomeBrowser.

Results: Low complexity microsatellites with $(A)_nB$ unite (exact phase) dominate in human and mouse genomes. *Alu*- and L-like elements overlap usually with $(A)_nB$-repeats, rather then with repeats of any other types. We have found all examples of such microsatellites flanking *Alu*-repeats. The youngest *Alu*'s overlap with microsatellites more frequently and microsatellites overlapping with younger *Alu*'s are more often exact.

Conclusion: As it is known, *Alu* integration is associated with L1-related endonuclease that cuts at AT-rich sequences [3]. It is also known that AT-rich sequences appear at the *Alu* ends and can give a birth to repeats of $B(A)_n$-type [4,5]. Thus *Alu*- and L-like elements of human and mouse respectively contribute substantially into genome composition of microsatellites with periods of 4-6. Particularly, contribution of *Alu* is decisive in the human genome. It appears that generation of *Alu*-adjacent microsatellites takes place simultaneously with *Alu* integration with subsequent degradation of both repeats.

Acknowledgements: This study was partially supported with Russian Fund of Basic Research Grant # 07-04-01584-a

References

1. S.S.Arcot (1995) *Alu* repeats: a source for the genesis of primate microsatellites. *Genomics*, **29**: 136-144.
2. M.A.Batzer, P.L. Deininger (2002) *Alu* repeats and human genomic diversity. *Nature*, **370**: 370-389.
3. J. Jurka (1997) Sequence patterns indicate an enzymatic involvement in integration of mammalia retroposons. *Proc. Natl. Acad. Sci. USA*, **94**: 1872-1877.
4. Nadir E. et al (1996) Microsatellite spreading in the human genome: evolutionary mechanisms and structural implications. *Proc. Natl. Acad. Sci. USA*, **93**: 6370-6475.
5. Toda Y. et al (2000) Characteristic sequence pattern in the 5- to 20-bp upstream region of primate *Alu* elements. *Mol. Evol.*, **50:** 232-237.

EVOLUTIONARY TRANSITION TO COMPLICATED DYNAMICS OF TWO-AGED POPULATION'S NUMBER

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

Motivation and Aim: Investigation of natural selection results in Mendelian panmictic isolated population of diploid organisms is the classic research task of mathematical population genetics [1]. In this task the homogeneous population is considered; i.e. it is suggested that different generations of this population are not overlapped. This suggestion is true for some natural populations, but such situation is rather exception than rule. Many biological species have age structure. In our issue we consider how the age structure presence may change the results of classic natural selection research.

On the other hand it is shown in [2] that reproductive potential and survivability growth is followed by the complication of number dynamics of two-aged population. It is natural to suggest that growth of individuals' fertility and of survivability in natural populations is the result of evolution under the natural selection.

Methods and Algorithms: The analytical and numeric investigation of natural selection model in two-aged population carried out:

$$\begin{cases} x_{n+1} = \overline{w}_n y_n, y_{n+1} = x_n(1-x_n) + cy_n \\ q_{n+1} = \dfrac{p_n(w_{AA}p_n + w_{Aa}(1-p_n))}{a_n}, p_{n+1} = \dfrac{x_n(1-x_n)q_n + cy_n p_n}{x_n(1-x_n) + cy_n} \end{cases}$$

here $\overline{w}_n = w_{AA}p_n^2 + 2w_{Aa}p_n(1-p_n) + w_{aa}(1-p_n)^2$ is reproductive potential of reproductive part of population (or average fitness of embryos); w_{ij} – genotype ij fitness, x_n – the number of junior age class in n-th reproduction season, and y_n - the number of reproductive part of population; q_n – frequency of allele A in junior age class and p_n – frequency of allele A in reproductive part of population.

Results and Conclusions: Our research shows, that growth of average fitness of embryos $\overline{w}$ is followed by the destabilization of the population number at first; than (if population is genetically polymorphic) occurs destabilization of genetic structure of population age groups. So it may be concluded, that transition from stable regimes of population numbers dynamics as well as of population genetic structure to fluctuations and chaos is result of the natural evolution in natural age-structured population.

References:

1. Fischer R.A. (1930) The Genetical Theory of natural selection. (Clarendon Press, Oxford).
2. Frisman E.Ya., Skaleckaya E.I. (1994) Strange attractors in simplest models of population numbers dynamics of biological populations, *Obozrenie Prikladnoj I Promyshlennoj Matematiki,* **6**: 988-1008.

NESTED ARC-ANNOTATED SEQUENCES AND STRONG FRAGMENTS

Furletova E.[1], Roytberg M.*[1], Starikovskaya T.[2]
[1]Institute of Mathematical Problems of Biology, Pushchino, Moscow Region, Russia
[2]Lomonosov Moscow State University, Moscow, Russia
*Corresponding author: e-mail: mroytberg@impb.psn.ru

Motivation and Aim: The nested arc-annotated sequences (NAAS) represent RNA secondary structures. Informally, a NAAS is a word in the alphabet {A, U, G, C} and a set of nested arcs connecting its letters. An *optimal structure* for a given word is a NAAS on the word having maximal possible number of arcs among the NAASs. A word is *strong*, if any of its optimal structures has an arc between the first and the last characters of the fragment. The run-time bound of a dynamic programming algorithm finding an optimal structure for a given word *w* can be improved by replacement of the number of all fragments of a word w (~ n^2, where *n* = *n(w)* is a length of the word *w*) by the number *F(w)* of strong fragments of the word *w* [1]. In [1] the authors claim that the average value F(w) is linear if we will consider only sequences meeting some physical restrictions. Our aim was to study the behavior of *F*(*n*) where *F*(*n*) is an average number of strong fragments of a random word w of length *n* (all letters of *w* are iid variables).

Methods and Algorithms: We have proved the following statement. Let *G(n)* be a number of strong words of length n; g*(n) = G(n)/4*^{n} . Then for all $n \geq 0$

$$F(n+1) = n * g(2) + (n-1) * g(3) + ...1 * g(n+1). \quad (1)$$

In order to understand the behavior of $F(n)$ we will study $g(n)$. If $g(n) \geq c > 0$ for all n, then the function $F(n)$ is quadratic.

Experiments and results: To investigate the average number of strong fragments of a random word we have performed Monte-Carlo computer experiments. For each $n \in \{1,\ldots, 1000\}$ we have created a set R[n] consisting of 10000 random sequences of length n. Besides this, we have calculated the experimental values of $F(n)$ and g(n) over the set. The experiments show that $g(n) \approx 0.02$ for all $n \geq 30$. Using (1) and information about g(n) we have approximated $F(n)$ by the formula $F^*(n) = 0.01n^2 + 0.64n - 3$ (n > 30). For all $n > 30$ we have $F(n) > F^*(n)$ and the difference $F(n) - F^*(n)$ grows monotonically for $n > 100$.

Conclusion: Formula (1) and computer experiments show that the average number *F(n)* of strong segments of a random sequence of length *n* growth ~n^2 .

References:

1. Wexler Y., Zilberstein C., Ziv-Ukelson M. A Study of Accessible Motifs and RNA Folding Complexity. Proceedings of RECOMB 2006: 473-487.

TRANSITIVE SUBSET SEEDS FOR PROTEIN ALIGNMENT

Furletova E.[1], Kucherov G.[2], Noe L.[2], Roytberg M.*[1], Tsitovich I.[3]
[1]Institute of Mathematical Problems of Biology, Pushchino, Moscow Region, Russia
[2] LIFL,Villeneuve d'Ascq, Cedex, France
[3] Institute for information transmission problems, Moscow, Russia
*Corresponding author: e-mail: mroytberg@impb.psn.ru

Motivation and Aim: In [1] we have introduced the class of subset seeds and have demonstrated their advantages for the DNA comparison. In case of protein seeds the number of possible seed letters (i.e. subsets of the set of amino acid pairs *AP*) is $\sim 2^{210}$ thus one has to start with the choice of seed alphabet. Here we consider hierarchical and non-hierarchical transitive alphabets. A letter *D* is *transitive* if *D* contains all pairs *(a,a)* and for all $p_1, p_2, p_3 \in AP$ if $(p_1, p_2), (p_2, p_3) \in D$ then $(p_1, p_3) \in D$. Transitive letters are in one-to-one correspondence with partitions of the amino acid alphabet A. The *transitive* alphabet is a set of transitive letters including *Match* that corresponds to the partition where all amino acids are separated. The *hierarchical* transitive alphabet (HTA) is a set of embedded letters. In a *non-hierarchical* transitive alphabet (NHTA) letters are not necessarily embedded. Transitive letters allow for a direct hashing scheme within the database search. Our aim is to find out if it is possible to design subset seeds with better or equal selectivity (for a given sensitivity) than more complicated BLASTP-like seeds or vector seeds.

Methods and Algorithms: To design the desired seeds we use the three stage procedure: (A) design the transitive alphabet (hierarchical or not); (B) reveal seeds with maximal selectivity for the given set of sensitivity values; (C) same as (B) for the multi seeds. To perform the step (A) every amino acid pair is associated with both background and foreground probabilities. The *background* distribution is the distribution of letters in the aligned independent sequences, and the *foreground* distribution corresponds to the really interesting alignments. Let $D = \{p_1, \ldots, p_k\} \subseteq AP$ be a subset seed letter; $b(D) = \sum_{i=1,k} b(p_i)$ and $f(S) = \sum_{i=1,k} f(p_i)$ be its background and foreground probabilities. To find a good HTA we start with $R_1 = Match$ and then recursively produce R_{k+1} from R_k according to the following algorithm. For all pairs of classes C_1, C_2 from the partition R_k we find $W(C_1, C_2) = f(Br(C_1, C_2))/ f(Br(C_1, C_2))$, where $Br(C_1, C_2) = \{(a,b) \in AP \mid a \in C_1, b \in C_2\}$. Then we unite C_1, C_2 having maximal value of $W(C_1, C_2)$ into one class of R_{k+1}. To find a good NHTA we use similar algorithm. To produce successors of a member of *k*-th generation we maintain not one but *P=50* partitions having best value of *W*. The *(k+1)*-th generation is formed as Q=200 best successors of k-th generation. Thus we have ~10000 and to pick out of NHTA we used different heuristics based on the two following ideas: (1) we prefer the letters with high likelihood ratio; (2) the alphabet should contain letters of different weights. Given a transitive alphabet, steps (B) and (C) can be performed with proper genetic algorithms.

Experiments and Results: We have shown that the transitive seeds of size 4 (both hierarchical and non-hierarchical) can outperform the efficiency of BLASTP-like seeds.

The seed selectivity was computed by the algorithm from [1] according to the Bernoully model with probabilities corresponding to the BLOSUM62 substitution matrix. The same result was also shown for the experiments with real testing sets (BAliBase, HOMESTRAD, etc.)

References:

1. G. Kucherov, L. Noe, M. Roytberg. A unifying framework for seed sensitivity and its application to subset seeds. Journal of Bioinformatics and Computational Biology 4 (2006) 553–570

MULTITASKING SOFTWARE SYSTEM FOR DNA ANALYSIS

Fursov M.*[1], Novikova O.[2]
[1] Novosibirsk Center of Information Technologies 'UniPro'.
e-mail: fmike@unipro.ru
[2] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
e-mail novikova@bionet.nsc.ru
* Corresponding author

Motivation and Aim: In the present paper we share our experience concerning the design and implementation of a scalable software system that is aimed to solve the problems in DNA analysis such as incompatibility and disconnection of different information sources, high specialization and lack of computational scalability of existing software tools, limited visualization interface, etc.

Variety of the incompatible information sources and methods constrains to the manual handling of inputs and outputs from different software tools. It is one of the main bottlenecks in the DNA analysis. The common problems are the necessity to use WEB applications to get a result, command line scripts and incompatible formats.

The solution is a software system that supports the set of information sources and computational methods and can be extended to support new ones.

Results, Conclusion and Availability: We designed and implemented an interactive visual software system called UniPro GenomeBrowser 2.0 [1] aimed to simplify work with DNA sequences, alignments and annotations. The supported data formats and the types of information sources, as well as analysis and visualization methods, can be added to the system without any changes to the existing code. The system has an internal resource manager and a task scheduler that make application run smooth on slow personal computers and utilize the resources of modern multicore systems.

Our results can be used by the architects of the similar systems as well as by researchers for immediate use and as an educational tool for computational methods in DNA analysis. The binary distribution of the system can be requested from the authors by email and will be demonstrated on a computer stand on BGRS'2008.

UniPro GenomeBrowser 2.0 web site: http://genome.unipro.ru/

NEW COMPUTATIONAL RESOURCES FOR THE STUDIES OF THE UBIQUITIN SYSTEM

Gainullin M.R.[1, 3]*, Chernorudskiy A.L.[1], Kovalyov V.A.[3], Eremin E.V.[2], Astashev M.E.[4], Garcia A.[1]

[1] Nizhny Novgorod State Medical Academy,
[2] Institute of Applied Physics RAS,
[3] Nizhny Novgorod State University, Nizhny Novgorod, Russia
[4] Institute of Cell Biophysics RAS, Pushchino, Russia
e-mail: biochem@gma.nnov.ru
* Corresponding author

Motivation and Aim: The ubiquitin system forms one of the most important pathways of cellular regulation. At present, the amount of experimental data concerning the ubiquitin system grows extensively. On the other hand, unraveling of the respective functional consequences is not so fast, and many of the ubiquitin-dependent regulatory mechanisms still remain enigmatic. We try to reduce this divergence by development of computational resources related to the ubiquitin system.

Results: The present research comprises several aims. To fill a general need for collecting and systematizing experimental data concerning ubiquitylation we have developed a specialized resource, UbiProt Database, a knowledgebase of ubiquitylated proteins (http://ubiprot.org.ru) [1]. The database contains retrievable information about more that 400 target proteins, as well as ubiquitylation features and related ubiquitylation/de-ubiquitylation machinery. The utility of UbiProt is also being extended for usage in proteomics tasks. To facilitate identification of ubiquitylated proteins by mass spectrometry, we have developed UbIdent, a new tool for virtual fragmentation of target proteins (included in UbiProt dataset or supplied by user) and calculation of the precise masses for peptides bearing modification. Currently we are going to represent the whole ubiquitin system in terms of the system biology. Formalized rules have been introduced for description of the ubiquitylation cascade and deubiquitylating enzymes, their interactions, upstream regulatory events and downstream metabolic outcomes. Respective resource termed «Ubiquitomix» is now under development using the BioUML workbench ((http://www.biouml.org) [2], which enables graphical representation and mathematical or computational modelling.

Conclusion: Computational resources presented here provide explanatory and predictive insights into the behaviour of the ubiquitin system and its role in cellular signalling.

Availability: All resources are freely available for non-commercial users.

References:

1. A.L.Chernorudskiy et al. (2007) UbiProt: a database of ubiquitylated proteins, *BMC Bioinformatics,* **8:** 126.
2. F.Kolpakov et al. (2007) CYCLONET--an integrated database on cell cycle regulation and carcinogenesis, *Nucleic Acids Res*, **35:** D550-D556.

WEB-TOOL FOR PROTEIN DESIGN BY THE ANIS-METHOD

Garkovenko A.V.*, Bogatova O.V., Kozmin Yu.P., Nekrasov A.N.
M.M. Shemyakin & Yu.A. Ovchinnikov Institute of Bioorganic Chemistry, RAS, Moscow, Russia. e-mail: garkovenko@gmail.com
* Corresponding author

Motivation and Aim: One of the important tasks is to obtain the minimal sized protein, retaining the ability to elicit essential biological function of the native protein. Truncated forms of protein with the single activity are applicable in biotechnology, pharmacology and medicine. The new approach to solve this task is suggested in this work;

Methods and Algorithms: Method of partitioning of the sequence is based on the new type of the approach of protein sequences using the informational units (IU) [1]. This approach allows using protein sequence alone to build the system of the hierarchically organized IDIC-sites. Full set of IDIC-sites is the protein information structure (IS). IDIC-sites are characterized by Increased Degree of Informational Coordination (IDIC) between amino acid residues and have high degree of structure stability. ANalysis of Information Structure method (ANIS-method) consists of the following steps:

-coding of the sequence using of the IU-approximation (informational units);
-making the informational density profile of the protein sequence;
-position identification of the IDIC-sites with the different hierarchical levels;
-making the IDIC-diagram (graphic presentation of the protein IS);

As the result the protein is presented as the system of hierarchically organized IDIC-sites of different lengths;

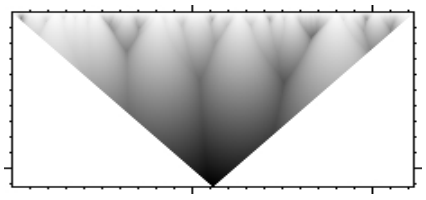
Results: Depending of the current task the researcher have the opportunity to choose the protein sequence fragment, truncation of which will minimally perturbate 3D structure of the rest part of the protein. Efficiency of the method was verified by obtaining the recombinant truncated forms of peroxiredoxin-6 [2] and interleukin-13 [3].

The WEB-service is organized as follows: after the free registration of the user, the amino acid sequence in one letter code maybe entered in the “SEQUENCE” window. After the computation the file with graphical representation of the protein informational structure is sent by e-mail to the user. Figures 1A and 1B represent IS of peroxyredoxin-6 and interleukin-13, obtained with the WEB-service. The patent application is filed for the developed method [4];

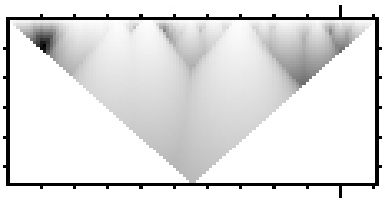
Conclusion: The WEB-service is set on the base of the ANIS-method, allowing to investigate hierarchical organization of IDIC-sites and being the new effective tool for the design of new recombinant proteins;

Availability: This web-service available on http://trees.anis.ibch.ru;

A



B



References:

1. A.N. Nekrasov (2004) Analysis of the information structure of protein sequences: a new method for analyzing the domain organization of proteins. *J Biomol. Struct. Dyn.,* **21**: 615-24.
2. A.N. Nekrasov et al. (2007) The novel approach to the protein design: active truncated forms of human 1-CYS peroxiredoxin. *J Biomol. Struct. Dyn.,* **24**: 455-62.
3. A.N. Nekrasov et al. (2008) The design of a novel IL-13 antagonist from the analysis of the informational structure, in press.
4. Patent application № 2007143253 (23.11.2007).

IN SILICO DESIGN OF PRIMER FOR 28 KDA ANTIGEN PRECURSOR PROTEIN OF *MYCOBACTERIUM LEPRAE*

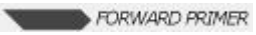
Gaur A.
Project Associate, Biotech park, Lucknow, India.
e-mail: www.adielixer@gmail.com

Motivation: Bioinformatics is growing, a field still existing in its aboriginal version demanding a vast scope to be exploited by the budding engineers and professionals working with life sciences having preliminary knowledge in informatics. Various tools and soft wares available on line along with vast database resource available at NCBI, EMBL, Swiss-Prot etc helps us bringing the biological approach of diagnosis to be operated at a remote computer.

Aim: Microbes, ever since, have served as one of the root causes to many diseases reaching up to deadly ones like AIDS, Leprosy, etc. It has always been Allopathic, Therapeutic or Ayurvedic (herbs), approach to combat the disease worldwide. The objective is to solve the problem using the tools of Bioinformatics and then apply it in vivo.

The project aims at designing an Insilico primer for **28 kDa antigen precursor protein of *M. Leprae***, which is a surface protein found on the membrane of leprae and recognizing which human body produces antibodies.

Methods and algorithms: The project is to build primer for the respective protein using different tools and soft wares available online (provided by various research institutes and universities like Genefisher, DNASIS-Max, uni-prot, NCBI, PDB, etc) and thereby doing a comparative study of the results. Depending upon the congruency of the results obtained with different soft wares, each using different Algorithm, my objective is to build the best forward and reverse primer for the protein. I have successfully developed the best primers Insilico. Once the primer has been build it can then be exploited in wet lab to produce the protein using diverse techniques of molecular biology and then finally it can be used to produce Antibodies (exclusively Monoclonal) for the protein and hence can be used to combat the disease in the sick patients of Leprosy.

Results: Best forward primer 

1. Matched in PRIMER3 and DNASIS-Max

Position- 67 to 87
Sequence- gtccatgcgcatatttcctt

2. Matched in WebPrimer and Genefisher

Position- 16 to 34
Sequence- gatgcaagctttcgaca

3. Matched in Genefisher and DNASIS-Max

Position- 245 to 273
Sequence- caccgatgatccaggta

Best reverse primer 

1. Matched in DNASIS-Max and PRIMER3

Position- 270 to 250
Sequence- caggccattacctggatcat

Conclusion: Project abstract once highlighted in medical community can bring out the possible approach to more remote technique and can help millions of people suffering from Leprosy; moreover, it would certainly highlight the role of Bio-informatics in solving the dilemma of medical science and putting it to a more real approach rather than considering it merely a computer one.

References:

1. V.M. Katoch**,** Int J Lepr Other Mycobact Dis. 2004 Jun;72(2):149-58.
2. Director clinical Divison, Central JALMA Institute for Leprosy, Tajganj, Agra 282 001, India.
3. P. K. Seth, Director, Bioinformatics division, Biotech park, India

SEQUENCE ANALYSIS OF THE GH70 FAMILY

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

Motivation and Aim: The GH70 family of glycoside hydrolases is composed by several dozens of proteins. Enzymatically characterized members of this family possess alternansucrase, dextransucrase, and reuteransucrase activities. Evolutionary relationships within the family are still unclear and became the purpose of the work.

Methods and Algorithms: Protein sequences were retrieved from the NCBI database. Multiple sequence alignment of 39 GH70 domains was made in BioEdit program (very similar and partial sequences were omitted). The phylogenetic trees were built using programs of PHYLIP package. Clusters of sequences with bootstrap support higher than 90% were considered as stable. Interfamily relationships were established by PSI-BLAST searches, using several most divergent representatives from the GH70 family.

Results: GH70 domain is almost always preceded by GH70N and followed by COG5263 domains. The majority of proteins do not have additional domains, but some of them contain DUF1542 or REUTERI. Several proteins include two GH70 or two COG5263 domains. *Exiguobacterium sibiricum* has an unusual circularly permutated GH70 domain. GH70N and REUTERI are new domain families with unknown functions. The pairwise comparison of GH70 domains showed that the sequence identity is higher than 30% for almost all pairs. Both Neighbor-Joining and Maximum Parsimony phylogenetic trees of the family have a similar topology. *E. sibiricum* GH70 sequence was used as outgroup. All other domains form two stable clusters. The smaller of them contains only *Lactobacillus reuteri* domains. Two main stable subclusters of sequences can be distinguished inside the second cluster. One of subclusters is composed by the rest part of *L. reuteri* sequences. The other one is formed by the majority of GH70 domains and includes a stable group of proteins from several *Streptococcus* species. Six the most divergent GH70 domains, selected based on the phylogenetic trees, were used as query for PSI-BLAST. Two iterations with each of them allowed us to obtain 3660 nonidentical protein sequences in total, representing GH13 and GH70 families. Particularly, members of 1, 2, 4-12, 14-17, 19, 20, 23, 24, 26-32, 35, and 36 subfamilies of the GH13 family were found.

Conclusion: Conservative domain structure of GH70 proteins and their representation almost only in the lactic acid bacteria suggest a recent appearance of GH70-encoding genes. Existence of several GH70-paralogues in many organisms and two GH70-domains in some proteins shows a significant role of duplications in the evolution of the family. Our data confirm close relationship of GH13 and GH70 families of glycoside hydrolases (they belong to GH-H clan). A similar circularly permutated structure of GH13 domains and *E. sibiricum* GH70 domain allows us to conclude that the latter resembles structure of GH70 domain ancestor.

We are thankful to Marco Carreras for help with automation of PSI-BLAST output analysis and to the Russian Foundation for Basic Research for a financial support (grant 06-04-49079-a).

MODELING OF THE LOOP ORGANIZATION OF EUKARYOTIC CHROMOSOMES *IN SILICO*

Glazkov M.V.
N.K. Koltsov Institute of Development Biology RAS, Moscow, Russia.
e-mail: MVGlazkov@yandex.ru

According to the loop-domain model, eukaryotic chromosomes are organized into the series of loop structures. The basements of these loops are anchored at the residual nuclear structures (nuclear matrix) (1).

This work is devoted to the question of the possible existence of loop organization of chromatin which does not require the anchoring of chromosomal DNA at the nuclear matrix. To this end, the computer modeling of the loop organization of chromosomal domains of genes of plants (2 genes), animals (4 genes), human (11 genes), and also short (human Alu 1, hamster Alu 2, mice B1) and long dispersed (LINE 1 of mice, of rats and of rabbit) repeats, the tandem repeats (satellite DNA from mouse centromeres and the γ-satellite of the mouse) has been carried out. The method suggests the search for «complementary» tracks of polypyrine and polypirimidine sequences capable to form three-stranded DNA structures (the H-form of DNA), that is accompanied by the looping of chromosomal DNA regions located between such tracks.

It is shown, that both gene-coding, and gene-not coding regions of eukaryotic chromosomes can be packed into the systems of small size loops (up to 5 kbp DNA) by the similar mechanism. Thus, chromosomal domains of genes (including located close to each other) are capable to be packaged into the isolated systems of loops that can be the structural basis of their independent expression. The character of loop organization of gene domains gives way to explain both the existence of some types of pseudo-genes and various duplications, deletions and inversions of some parts of genes, including those leading to hereditary diseases.

Results of modeling are in good conformity with the data of electron microscopic research of the loop organization of eukaryotic chromosomes, the study of fragments of the chromosomal DNA isolated from core of rosette-like structures of interphase chromosomes, modeling of the loop organization of human α_1-antytrypsine gene in vitro obtained earlier in our laboratory (2,3,4), and also the well-known facts about more compact packaging of the chromosome regions containing satellite DNA and the preference association of blocks of satellite DNA with the nuclear envelope.

References:

1. S.M.Gassser, U.K.Laemmli (1987) A glimps at chromosomal order, *Trends in Genetics*, **3:** 16-22.
2. M.V.Glazkov (1989) Ultrastructure of somatic and meiotic nucleoids, *Electron Micscopic Rev*, **2:** 197-229.
3. M.V.Glazkov et al. (1994) DNA sequences isolated from protein cores of rosette like structures (elementary chromomeres) of mouse interphase chromosomes, *Russian J. of Genetics*, **30:** 993-1000.
4. O.S.Chudinov, M.V.Glazkov (2002) In vitro formation of triple-stranded DNA structures in the human α_1-antitrypsin gene, *Russian J. of Mol Biology*, **36:** 123-128.

DEVELOPMENT OF NEW GENETIC METHODS FOR PREDICTIVE TESTING OF MULTIFACTORIAL DISEASES AND MAXIMUM PROLONGATION OF THE HUMAN ACTIVE LIFE (ANALYSIS OF 16 GENES)

Glotov O.S.[1, 3*], Glotov A.S.[1, 2], Demin G.S.[1], Potulova S.V. [1, 2], Moskalenko M.V.[2], Shved N.Y.[1], Vakharlovsky V.G.[1], Ivashchenko T.E.[1], Baranov V.S.[1, 2]

[1] Ott's Institute of Obstetrics and Gynecology RAMS, St-Petersburg, Russia

[2] St-Petersburg State University, St-Petersburg, Russia

[3] "The Eternal Youth" Foundation, St-Petersburg, Russia

e-mail:olglotov@mail.ru

* Corresponding author

Motivation and Aim: It is well-known that diagnostics on early stages and prophylaxis of different diseases is one of the most actual problems of modern medicine. Patient usually consults a doctor on the stage when disease is already in progress. At the same time the cause of pathological process often remains unclear. In this case intensive treatment can improve state of health of patient only for sometime and makes him dependent on symptomatic drugs. Diagnostic of predisposition to the multifactorial pathologies nowadays becomes an important tool that is necessary for solution of the predictive medicine problems. The basic directions of these researches are connected to genes of cancerogenesis and cardiovascular diseases.

Methods and Algorithms: Using pharmacy biochip specially constructed for population studies, 13 polymorphisms of 7 genes: *CYP1A1*(C4887A, A4889G, T6235C), *CYP2D6*(G1934A, DelA2637), *CYP2C9*(C430T, A1075C), *GSTM1*(del), *GSTT1*(del), *NAT2*(C481T, G590A, G857A), *CYP2C19*(G681A) and using RFLP method 9 polymorphisms of genes: *AGT*(M235T), *ACE*(I/D), *AGTR1*(A1166C), *PAI1*(4G/5G), *GPIIIa*(C1565T), *MTHFR*(C677T), *NOS3*(4/5), *IGF-1*(CA repeats), *PGC-1*(Gly482Ser) were investigated in 3 age-specific groups from North-West Region of Russia.

Results: The frequencies of same genotypes and alleles of *CYP2C9, GSTM1, GSTT1, NAT2, AGT, ACE, AGTR1, PAI1, MTHFR, NOS3, IGF-1, PGC-1 genes* were different between studied groups.

Conclusion: In our investigation was demonstrated that people who have certain genotypes of studied genes for men and for women have some metabolic advantages for their longer survival. Further, it is necessary to perform studies on various groups of different age, taking into account meta-analysis data to estimate the role of age-regulating genes and multifactorial diseases in aging.

CLASSICAL ATTENUATION REGULATION MODEL

Glotova I. *[1], Lyubetsky V. [2]
[1] Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Moscow, Russia
[2] Institute for Information Transmission Problems RAS (Kharkevich institute), Moscow, Russia
e-mail: igbox@mail.ru
* Corresponding author

Motivation and Aim: Modeling regulation processes of gene expression in bacteria is a problem of great interest, in particular, it can be used as an alternative tool for the prediction of new regulatory sites. We have improved our model of classical attenuation regulation [1] and implemented it on the leader regions of amino acid biosynthetic operons in the complete bacterial genomes accessible in GenBank.

Methods and Algorithms: The performance of our model improved due to two basic changes: (1) folding paths with pseudoknots as well as (2) the influence of RNA triplex formation are currently considered in the process of secondary structure formation in the leader region. The procedure of secondary structure energy calculation was intensively analyzed.

Results: The improved model was implemented under parameters with the same fixed value on the bacterial operons given below. The small part of the results presented refer to gamma-proteobacteria – the most studied ones. "A" stands for the cases when classical attenuation regulation was predicted through bioinformatics methods and/or experimentally confirmed, "P" - for the cases when such regulation was predicted by our model.

Gamma-proteobacteria	*hisG*	*pheA*	*pheS*	*trpE*	*thrA*	*leuA*	*ilvA*	*ilvC*	*ilvB*	*ilvG*	*ilvI*
Escherichia coli	A, P	A	A, P	A, P	A, P	A			A, P	A	
Salmonella typhi	A, P	A	A, P	A	A, P	A			A	A, P	
Klebsiella pneumoniae	A, P	A	A, P	A, P	A, P	A			A, P	A	
Erwinia carotovora		A	A	A	A, P	A			A, P	A	
Yersinia pestis	A, P		A, P	A, P	A				A, P	A	
Haemophylus influenzae	A, P				A, P	A				A, P	
Pasterella multocida					A, P	A				A	
Vibrio cholerae	A	A, P		A, P	A	A, P				A, P	
Vibrio vulnificus				A, P							
Vibrio parahaemolyticus				A, P							
Shewanella oneidensis		A, P		A	A, P	A				A, P	
Idiomarina loihiensis L2TR					P						
Pseudoalteromonas haloplanktis TAC125					P						
Pseudomonas putida										A, P	
Pseudomonas syringae										A, P	

Conclusion: An improved model of classical attenuation regulation has been proposed. Its implementation allowed us to predict the existence of such regulation in the leader regions in considerable number of proteobacteria and actinobacteria.

Availability: http://lab6.iitp.ru/rnamodel

References:

1. V.A. Lyubetsky, S.A. Pirogov, L.I. Rubanov, A.V. Seliverstov, "Modeling classic attenuation regulation of gene expression in bacteria", Journal of Bioinformatics and Computational Biology, vol.5, no.1, 2007, pp. 155-180.

APPLICATION OF HIDDEN MARKOV MODELS FOR THE SEARCH OF TRANSCRIPTION FACTOR BINDING SITES

Golda R.Ya.
National University of "Kyiv-Mohyla Academy", Kyiv, Ukraine.
e-mail: tx_hv@ukr.net

Motivation and aim: To find the genes of primary cellular response to external factor, transcription factor binding site (TFBS) search can be performed. Finding TFBS by consensus or by position-specific scoring matrices has insufficient accuracy and it is necessary to apply additional filters. Therefore, we proposed to use hidden Markov models (HMM), which allow to preserve more information than matrices.

Methods and algorithms: To conduct the search, we formed two-dimensional Markov model with 4xN size, where N is the length of TFBS. To calculate probabilities in Markov chain links, three-dimensional matrix is calculated from known TFBS sequences. Also, we inserted into the model position-specific scoring algorithm as a self-linking probability. Cut-off was calculated using known sequences, assuming similar score distribution for unspecified TFBS sequences.

Results: This method was applied for finding ISRE (Interferon-Stimulated Responsible Element). We used ISRE position-frequency matrix from TRANSFAC 7.0 Public and 8 known ISRE sequences. Promoters are defined as 800 bp upstream the transcription start site plus the 5`UTR. Using this method, we found 331 genes with putative ISREs. Functional analisys was conducted using Gene Ontology Tree Machine. As a result of functional analisys, 13 categories were identified. In biological process: regulation of signal transduction, positive regulation of signal transduction, protein biosynthesis, immune effector process, B cell mediated immunity, immune response, adaptive immune response, adaptive immune response (sensu Gnathostomata), response to virus; in molecular function: enzyme activator activity, phospholipid transporter activity; in cellular component: MHC protein complex, MHC class I protein complex. In the gene list we have genes of immune response, cytokines precursors, and cytokine regulative genes, growth factors, and apoptosis genes. Several genes are already annotated, as interferon-alpha induced genes.

Conclusion: We developed hidden Markov models-based method for finding transcription factor binding sites. Results of ISRE search (quantity of genes and functional categories) prove that the proposed method is applicable.

Availability: Finding TFBS using hidden Markov models is possible using COTRASIF: http://biomed.org.ua/COTRASIF/

ALLELIC VARIATIONS AT THE VRN1 GENE PROMOTER SEQUENCES RESPONSIBLE FOR VERNALIZATION IN WILD AND DOMESTICATED WHEATS

Golovnina K.A.*, Kondratenko E.Y., Blinov A.G., Goncharov N.P.
Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
e-mail: ksu@bionet.nsc.ru
* Corresponding author

Motivation and Aim: Understanding the process of crops domestication gives a clue for the mechanism of morphological diversification during evolution. Recently, it has been demonstrated that transcriptional regulators act as switches between discrete developmental programs, encouraging the view that novel morphological differences may arise from changes in this class of genes. Vernalization, the requirement of a long exposure to low temperatures to induce flowering, is an essential adaptation of plants to cold winters. It has been shown that the deletions in the promoter region of gene the *VRN1* gene from cultured diploid wheat *T. monococcum* were associated with spring growth habit. The aim of the present work was to study allelic variability in *VRN1* promoter region of some polyploid and wild diploid species together with goat grasses known as the genomes donors for polyploid wheats.

Methods and Algorithms: Accessions of the wheat species and goat grass were obtained from the different collections, including Russia, Japan, USA, Syria, and the Netherlands. In the present work all used accessions were grown and their generic definitions were performed. Experimental approaches included total DNA isolation, development of genome-specific primers, PCR amplification, and sequencing.

Results: Most of the wild Triticeae have a winter growth habit, suggesting that the recessive *vrn1* allele is the ancestral character. In contrast, the majority of cultured polyploid wheat species are spring due to at least one dominant *Vrn1* allele that could be inheriting or appeared independently as a result of the intensive selection on this trait during domestication process. In the present work we investigated the unique collection of the wild diploid species *T. urartu* and *T. boeoticum* together with *T. monococcum*, goat grasses, durum and common wheats with the different dominant *Vrn-1* alleles. Deletions of the different size (22, 36, 55 bp) covered specific area of MADS-box were observed in *VRN1* promoter of the spring *T. monococcum* (K18105), *T. boeoticum* (K40117, K40118, PI427328, K20741, IG45296), in one allele of *T. urartu* (PI428276), and A genome of durum wheat. Data obtained in this work illustrated the extension of deletion region in *VRN1* promoter of the polyploid species during their evolution. Specific indels and nucleotide substitutions in A, B and D genomes sequences of investigated region were observed. At the same time, A genome sequences variability of the diploid species included haplotypes specific to all three genomes. An absence of damages in *VRN1* promoters of spring common wheat samples found out in winter populations shew an existence of other vernalization mechanisms that may appear independently in these species during domestication. Goat grasses were supposed as B, G, and D genomes donors of polyploid wheats and association of their spring grow habit with deletions in *VRN1* promoter region was not observed so as in B and D genomes of polyploids.

Conclusion: This work is an initial insight into complicated analysis of evolution process took place in agricultural genes during evolution and involved transcription factors.

ON STABILITY OF CYCLES IN GENE NETWORKS MODELS

Golubyatnikov V.P.*, Gaidov Yu.A.**
* Sobolev Institute of Mathematics, SB RAS, Novosibirsk, Russia
** Novosibirsk State Pedagogical University, Novosibirsk, Russia
e-mail: glbtn@math.nsc.ru
* Corresponding author

Motivation and Aim: We continue our studies (see [1]) of phase portraits of nonlinear dynamical systems modeling asymmetric gene networks, in particular we are interested in behavior of their cycles in the cases when the bifurcation theory can not be applied. Our main goal now is to find conditions of stability of "non-bifurcation" periodic trajectories.

Methods and Algorithms: Our approach to these stability problems is based on theory of ordinary differential equations and on geometrical and topological methods related to the Conely index theory, see for example [2], [3].

Results: Consider dynamical system which represents a kinetic model of negative feedback regulation of 3-dimensional gene network functioning:

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

$x_i \geq 0$, *i = 1,2,3;* by convention $x_0 \equiv x_3$. Here the functions $f_i(x_{i-1})$ describe initiation of mRNA and/or proteins synthesis, they are positive, smooth, and monotonically decreasing. This system has a unique stationary point M_* , its topological index is -1. In [1] we have already described non-convex bounded invariant domain of this system.

Theorem. *If* M_* *is not stable and* $\max\limits_{i,x} |\frac{df_i}{dt}(x)| < 7 \cdot \min\limits_{i,x} |\frac{df_i}{dt}(x)|$ *then the dynamical system (1) has at least one orbitally stable cycle in the invariant domain.*

Actually, this stability estimate is not sharp, nevertheless it gives mathematical explanation of some numerical experiments with periodic trajectories and other attractors which appear in different models of gene networks.

Conclusion and future work: Our present task is to obtain similar stability results for higher-dimensional dynamical systems and for other types of gene networks models. One of the most significant properties of these more complicated models is their multistability. We plan to find and to classify the stable and periodic regimes of functioning of corresponding gene networks using Floer homology constructions described in [3], [4].

The work was supported by leading scientific schools grant 8526.2006.1 and by SB RAS, project 46. The authors are indebted to Ya.M.Eliashberg for helpful discussions.

References:

1. Yu.A.Gaidov, V.P.Golubyatnikov (2007) On some nonlinear dynamical systems modeling asymmetric gene networks, *Bulletin of Novosibirsk State University,* **7:** 23-32.
2. R.A.Smith (1987) Orbital stability for ordinary differential equations. *Journal of Diff. Equations,* **69**: 265-287.
3. D.A.Salamon (1990) Morse theory, the Conley index and Floer homology, *Bull.Lond.Math.Soc.,* **22:** 113-140.
4. T.Kaczynsky, K.Mischaikov, M.Mrozek (2004) *Computational homology*, (Springer).

TOPOLOGICAL INDEX OF A MODEL OF p53 DYNAMICS TRIGGERED BY DNA DAMAGE

Golubyatnikov V.P.*, Mjolsness E.**
* Sobolev Institute of Mathematics, SB RAS, Novosibirsk, Russia
** Department of Computer Science, University of California, Irvine, USA
e-mail: glbtn@math.nsc.ru
* Corresponding author

Motivation and Aim: We study the phase portrait of a system of nonlinear differential equations of dynamics proposed in [1] as a model of oscillations in the p53-Mdm2 DNA damage repair network [2]. Detailed analysis of the oscillating regimes of similar regulatory networks models with multistability properties is an important task both from the mathematical and from the biological points of view.

Methods and Algorithms: Our description of this nonlinear dynamical system is based on topological methods which we have elaborated in previous studies ([3]).

Results: The following dynamical system represents a model of oscillations in the p53-Mdm2 network (see [1]):

$$\frac{dx}{dt}=\alpha_0 z+\frac{\alpha_1\cdot x^6}{k_1+x^6}-\gamma_1 xy-\gamma_2 x\ ;\qquad \frac{dy}{dt}=\alpha_2+\frac{\alpha_3\cdot x^4}{k_2+x^4}-\gamma_3 y\ ;$$
$$\frac{dz}{dt}=\frac{\alpha_{1s}\cdot z\cdot(B-z)}{k_{1s}+B-z}-\frac{\alpha_{2s}\cdot z}{(k_{0d}+D)(k_{2s}+z)}\ ;\ \frac{dD}{dt}=-\alpha_d\cdot D\cdot zx\ . \qquad (1)$$

Here the variables x and y denote concentrations of p53 and Mdm2 respectively, z=*Atm-P* is the switch variable, and *D(t)* describes the DNA damage. In our numerical experiments we have taken the values of all parameters close to [1]. We construct a domain $Q(t)=[0,C_1]\times[0,C_2]\times[\varepsilon,B]\subset R^3_+(x,y,z)$ such that for small t all trajectories of (1) enter *Q(t)*. Hence for small values of t the sum of all indices of singular points in *Q(t)* equals -1. For large t and any $\varepsilon\geq 0$ this index vanishes. This shows the direct relation of the topological index of the repair system (1) and the presence of DNA damage.

Conclusion: Our analysis of the p53-Mdm2 network dynamics is based on a general approach connected with decomposition of high-dimensional models to lower-dimensional ones. Similar decompositions can be used in mathematical studies of more complicated regulatory networks.

The work was partially supported by the leading scientific schools grant 8526.2006.1, by SB RAS, project 46, and by US NIH grant P50-GM76516.

References:

1. V.Chikarmane, A.Ray, H.M.Sauro, A.Nadim (2007) A model for p53 triggered by DNA damage. *SIAM J. Appl. Dynamic Systems,* **6:** 61-76.
2. G.Lahav, N.Rosenfeld, A.Sigal, N.Geva-Zatorsky, A.J.Levine, M.Elowitz, U.Alon (2004) Dynamics of the p53-Mdm2 feedback loop in individual cell. *Natural Genetics,* **36:** 147-150.
3. Yu.A.Gaidov, V.P.Golubyatnikov (2007) On some nonlinear dynamical systems modeling asymmetric gene networks, *Bulletin of Novosibirsk State University,* **7:** 23-32.

INFERRING OPTIMAL SCENARIO OF GENE EVOLUTION ALONG A SPECIES TREE

Gorbunov K.Yu.*, Kanovei V.G., Lyubetsky V.A.
Institute for Information Transmission Problems of the Russian Academy of Sciences (Kharkevich Institute), Moscow, Russia
e-mail: gorbunov@iitp.ru
* Corresponding author

Motivation and Aim: Usual definition of mapping of a gene tree onto a species tree allows only indirect descriptions of horizontal gene transfer (HGT) events. A new approach ("inner tree mapping") is explicitly based on minimizing the amount of various evolutionary events.

Methods and Algorithms: Gene tree G and species tree S are given, with each leaf in G assigned a gene from a leaf species in S. Within each tree a "formal root" is placed and connected to the "old" root by a "root edge". Edges in S are subdivided into "temporal slices". Lets visualize and call edges in S as "tubes". Now consider positioning of all edges of tree G' inside the tubes; we refer to G' as the "inner" tree. Its formal root should be within the root tube of S and have edges going within the tubes toward the leaves. The edge of inner tree G' bifurcates together with the tube (speciation event) but can branch off further within the tube (duplication event), terminate within the tube (loss event) or leave one tube and penetrate another (with or without prior bifurcation, i.e. HGT event with retaining gene copy in the host or not). HGT event can occur only when the donor and recipient tubes are not genealogically descendant, or, more generally, when the two tubes belong to the same temporal slice; the latter requirement is dropped if slices are not defined. If the edge arrives in leaf v of tree S, its terminus is assigned a gene from v. Upon construction, inner tree G' is to meet the following requirements:

a) be isomorphic with given tree G, including leaves-assigned genes, after removal of all terminated (due to loss events) edges;

b) the penalty of the inner tree defined as a weighted sum of all events except for speciation is minimized, e.g. weight of loss is 2, of duplication – 3, of HGT with retaining gene copy – 11 and without retaining – 13.

Such inner tree G' and the corresponding mapping of G to S are efficiently found with the novel fast algorithm.

Results: The algorithm was thoroughly tested on artificial and biological data. For instance, with artificially defined G=(((*a*,*b*),((*c*,*d*),(*e*,*f*))),(*g*,*h*)) and S=(((*a*,*b*),(*c*,*d*)),((*e*,*f*),(*g*,*h*))), the algorithm produced a mapping with one HGT event (with retaining the host gene copy) from edge {*c*,*d*} into edge {*e*,*f*} and one loss event on edge {*e*,*f*}. Here an edge is denoted by the set of leaves contained in its corresponding subtree.

Conclusion: With data from [1] the algorithm outputs the same HGTs as in [1], as well as new biologically reasonable predictions.

References:

1. V. Lyubetsky et al (2005) Algorithms to reconstruct evolutionary events at molecular level and infer species phylogeny, In: *Bioinformatics of Genome Regulation and Structure* II, p. 189-204.

RELATIVE EFFECTS OF MUTABILITY AND SELECTION ON SINGLE NUCLEOTIDE POLYMORPHISMS IN TRANSCRIBED REGIONS OF THE HUMAN GENOME

Gorlov I.P. *, Gorlova O.Y., Amos Ch.I.
Department of Epidemiology, The University of Texas M. D. Anderson Cancer Center, Houston, Texas 77030
* Corresponding author: e-maik: ipgorlov@mdanderson.org

Motivation: Single nucleotide polymorphisms (SNPs) are the most common type of genetic variation in humans. However, the factors that affect SNP density are poorly understood. The goal of this study was to estimate the relative effects of mutability and selection on SNP density in transcribed regions of human genes. It is important for prediction of the regions that harbor functional polymorphisms. The goal of this analysis was to estimate the relative effects of mutability and selection on SNP density. We used data on ancestral and derived alleles from the dbSNP database (http://www.ncbi.nlm.nih.gov/sites/entrez?db=snp&cmd=search&term) and the Haplotter database (http://hg-wen.uchicago.edu/selection/haplotter.htm).

Methods and Algorithms: SNPs were stratified into five functional categories: (i) 5′ untranslated region (UTR) SNPs, (ii) 3′UTR SNPs, (iii) synonymous SNPs, (iv) SNPs producing conservative missense mutations, and (v) SNPs producing radical missense substitutions. Missense substitutions were further stratified as radical or conservative. The number of potential sites in the human genome for different types of SNPs were estimated based on codon composition (coding regions) and nucleotide composition (UTR regions). Only frequency-validated SNPs were used. Each of the SNP categories was further subdivided into nine mutational categories on the basis of the single-nucleotide substitution type. Thus, 45 functional/mutational categories were analyzed. The relative mutation rate in each mutational category was estimated on the basis of published data. The proportion of segregating sites (PSSs) for each functional/mutational category was estimated by dividing the observed number of SNPs by the number of potential sites in the genome for a given functional/mutational category. We used analyses of variance and covariance to estimate the relative effects of selection (functional category) and mutability (relative mutation rate) on the PSSs.

Results and Conclusions: By analyzing each functional group separately, we found significant positive correlations between PSSs and relative mutation rates (Spearman's correlation coefficient, at least r=0.95, df=9, $P<0.001$). We adjusted the PSSs for the mutation rate and found that the functional category had a significant effect on SNP density (F=3.3, df=4, $P=0.03$), suggesting that selection also affects SNP density in transcribed regions of the genome. We found that approximately 90% of variation in PSS was due to variation in the mutation rate and approximately 10% was due to selection, suggesting that the probability that a site located in a transcribed region of a gene is polymorphic mostly depends on the mutability of the site. Our results show that the adjusted PSSs in the 5′UTRs are similar to, or even lower than, the PSSs for radical missense mutations. This result suggests that 5′UTRs are under selection purification pressure as strong as that affecting radical missense mutations.

CLIDAPA: A NEW APPROACH FOR COMBINING CLINICAL DATA WITH GENES EXPRESSIONS

Guerra L.[1], González S.[1], Robles V.[1], Peña J.M.[1], Famili F.[2]
[1] Department of Computer Architecture and Technology, Universidad Politécnica de Madrid, Spain
[2] NRC Institute for Information Technology, Ottawa, Canada
e-mail: lguerra@laurel.datsi.fi.upm.es, sgonzalez@fi.upm.es, vrobles@fi.upm.es, jmpena@fi.upm.es, fazel.famili@nrc-cnrc.gc.ca

Motivation and Aim: Traditionally, clinical data have been used as the only source of information to diagnose diseases. Nowadays, other kinds of information, such as various forms of omics data (e.g. DNA microarrays), are taken into account to improve diagnosis and even prognosis in many diseases. This paper proposes a new approach for efficiently combining both sources of information, namely clinical data and genes expressions, in order to further improve estimations.

Methods and Algorithms: Clinical data and genes expressions are combined using a new algorithm, named CliDaPa. In this approach, patients are firstly divided into different clusters (represented as a decision tree) depending on their clinical information. CliDaPa is a greedy algorithm, developed with Java, Knime and Weka, which allow us to compute these clusters, depending on the percentage of successful classifieds. Thus, we obtain different groups of patients with similar behaviors. Each individual group can be studied and classified separately, using only gene expression data, with different supervised classification methods, like decision trees, Bayesian networks or lazy induction learning.

Results: To test our method, the well-known Van't Veer dataset on Breast Cancer has been used. For the proposed approach, internal (0.632 Bootstrap) and external validations (hold-out) have been carried out. Results have shown an approximate 15% improvements on accuracy in the internal validation and 5% of accuracy in the external validation in comparison with the traditional uses of clinical data and genes expression data separately.

Conclusions: Our new approach efficiently combines clinical and gene expression data, outperforming the traditional use of these sources of information. Thus, CliDaPa algorithm gets the fulfillment of the proposed objectives.

Availability: CliDaPa is available for the research community, for further information, please, contact the authors.

References:

1. Knime, konstanz information miner. (2007) URL www.knime.org.
2. J. Brenton. (2005) Molecular classification and molecular forecasting of breast cancer: ready for clinical application? J. Clin. Oncol., 23:7350–7360
3. E. R. Dougherty and U. Braga-Neto (2004) Is cross-validation valid for small-sample microarray classification? Bioinformatics, 20:2465–2472
4. B. Efron and R. Tibshirani (1997) Improvements on cross-validation: The 0.632 bootstrap method. American Statistical Assotiation, 92:548–560,
5. E. Fix and J.L.Hodges Jr. (1951) Discriminatory analysis, nonparametric discrimination.
6. Bloom HJG and Richardson WW. (1957) Histological grading and prognosis in breast cancer. Br J Cancer, II: 359–377.
7. P. Larrañaga, B. Calvo, R. Santana, C. Bielza, J. Galdiano, I. Inza, J. A. Lozano, R. Armañanzas, G. Santafé, A. Pérez and V. Robles (2006) Machine learning in bioinformatics. Briefing in Bioinformatics.
8. Yijun Sun, Seteve Goodison, Jian Li, Li Liu, and W. Farmerie (2007) Improved breast cancer prognosis through the combination of clinical and genetic markers. Bioinformatics, 23:30–37.
9. L. J. van 't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. Hart, M. Mao, H. L. Peterse, K. van der Koõy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Frieñd (2002) Gene expression profiling predicts clinical outcome of breast cancer. Nature, 415(6871):530–536, January.
10. T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downning, M.A. Caligiuri, C.D. Bloomfield, E.S. Lander (1999) Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring.
11. S. Paoli, G. Jurman, D. Albanese, S. Merler, and C. Furlanello, Semisupervised Proling of Gene Expressions and Clinical Data, ITC-irst - Trento, Italy
12. Nathalie L.M.M. Pochet, Frizo A.L. Janssens, Frank De Smet, Kathleen Marchal, Ignace B. Vergote, Johan A.K. Suykens and Bart L.R. De Moor, M@CBETH: Optimizing Clinical Microarray Classification, Department of Electrical Engineering ESAT-SCD, Leuven-Heverlee, Belgium

A MOLECULAR-GENETIC SYSTEM OF DEVELOPMENT: FUNCTIONAL DYNAMICS AND MOLECULAR EVOLUTION OF HH-, DPP- AND WG- SIGNAL CASCADES

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

Motivation and Aim: We compare parametric robustness of the Hh-, Dpp- and Wg- signal cascade mathematical models and molecular evolution of the Hh-, Dpp- and Wg- cascade proteins (genes).

Methods and Algorithms: Parametric robustness of the signal cascade mathematical models was taken from published data (Table 1). Gene evolution modes were analyzed in two steps [1; 2]: (1) searching of protein sites for which there was positive selection at the codon level in corresponding gene; (2) searching for branches of the phylogenic tree with positive selection by the computer test based on modeling of protein evolution [1]. Protein regions in which positive selection was identified at the first step were analyzed at the second step.

Results and Conclusion: Positive selection of the Hh-, Dpp- and Wg- signal cascade genes and the kinetic parameters of the Hh-, Dpp- and Wg- signal cascade models whose change produces the greatest shift in the cascade dynamics are related (Table 1). Therefore it may be assumed that the genes determining the hyper-response parameters had very important role in evolution. Even small changes in their function might have resulted in great changes in the function of the entire network and, hence, these genes could serve as candidates for "the within" sources of compensatory shift produced by mere point mutations.

Table 1. Examples of relations between gene evolution modes, divergence of Bilateria taxonomic groups, and hyper-responsive kinetic parameters.

Protein (gene) name	Network response corresponding to change in kinetic parameters (+ - hyper-response; - - inertness	Functional protein group	Events of positive selections related with divergence of Bilateria taxonomic groups [1; 2]
Hh	+*	Developmental	+
Dpp	+**	Developmental	+
Smo	+*	Developmental	+
Tkv	+**	Developmental	+
Fz2	+***	Developmental	+
Fz	+***	Developmental	+
PKA	-*	Housekeeping	-
Slmb	-*	Housekeeping	-
Ci	+*	Developmental	+
Mad	+**	Developmental	+

Designations: models for
*- the Hh-cascade [Lai, *et al.*, (2004) *Bio-phys. J.*, **86**: 2748-2757; Gunbin, *et al.*, (2007) *J. Bioinform. Comput. Biol.*, **5**: 491-506.];
- the Dpp-cascade [Shimmi, *et al.,* (2005) *Cell*, **120: 873-886; Umulis, *et al.* (2006) *Proc. Natl. Acad. Sci. U.S.A.*, **103**: 11613-11618];
*** - the Wg-cascade [Buceta, *et al.* (2007) *PLoS* **2**:e602; Amonlirdviman, *et al.* (2005) *Science* **307**: 423-426]

Availability: The detailed results are available from the authors upon request.

References:

1. K.V. Gunbin, et al. (2007) The evolution of the Hh-signaling pathway genes: a computer-assisted study, *In Silico Biol.*, **7**: 333-354.
2. K.V. Gunbin, et al. (2007) Aromorphoses and the adaptive molecular evolution, *Informational Herald of the VOGiS,* **11**: 373-400. (In russian)

THE MOLECULAR EVOLUTION OF THE PYROCOCCUS GENOMES: A COMPUTER-ASSISTED STUDY

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

Motivation and Aim: Our aim was to reveal genes that can be responsible for adaptation to high pressure environment conditions in hyper-thermophilic archaea genus *Pyrococcus*: *P. furiosus* grows at the shallow water; *P. horikoshii* grows at 1400 m depth, *P. abyssi* grows at 2000 depth. We used *Thermococcus kodakaraensis* as outgroup to reconstruct the *Pyrococcus* phylogeny.

Methods and Algorithms: Groups of orthologous proteins from *Pyrococcus* and *Thermococcus* proteomes [GenBank, release 160] was made by BLASTClust [1]. Multiple alignments of ortholog proteins were made by MAFFT 6.240 [2]; there were the sources for constructing aligned cDNA sequences. *Pyrococcus* phylogeny was reconstructed from the concatenation of multiple alignments of amino acid sequences by TREE-PUZZLE 5.2 [3]. Detection of gene and protein regions evolving under positive selection were made using Rate4Site 2.01 [4] and SNAP.pl correspondingly. Evolution of G+C contents was investigated using nhPhyML [5].

Results and Conclusion: We analyzed set of 1165 proteins that are presented in all three *Pyrococcus* genomes. Significantly higher rate of the Ka/Ks rates between *P. furiosus* and other *Pyrococcus* species were found for 11 genes: 9 genes were related to membrane functions, amino acid metabolism, DNA and protein synthesis (Table 1); 2 genes were uncharacterized. The results suggest that these functional properties of proteins can be particularly important for adaptation to the high pressure environment conditions.

Table 1. Functions of genes under positive selection pressure

COG #	COG functional class	Gene name
COG0438	Cell wall/membrane/envelope biogenesis	Glycosyltransferase
COG0733	General function prediction only	Na+-dependent transporters of the SNF family
COG1175	Carbohydrate transport and metabolism	ABC-type sugar transport systems, permease
COG1244	General function prediction only	Predicted Fe-S oxidoreductase
COG0626	Amino acid transport and metabolism	Cystathionine β-lyases/γ-synthases
COG0532	Translation, ribosomal structure	Translation initiation factor 2
COG1599	Replication, recombination and repair	ssDNA-binding replication protein A

Availability: The detailed results are available from the authors upon request.
The work was supported by SB RAS integration project #49.

References:

1. S.F. Altschul et al. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215**: 403-410.
2. K. Katoh et al. (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.,* **33**: 511-518.
3. H.A. Schmidt et al. (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics.* **18**:502-504.
4. I. Mayrose et al. (2004) Comparison of site-specific rate-inference methods: Bayesian methods are superior. *Mol. Biol. Evol.* **21**: 1781-1791.
5. B. Boussau and M. Gouy (2006) Efficient Likelihood Computations with Nonreversible Models of Evolution. *Syst. Biol.*, **55**: 756-768.

WHY TATA-BOX HIDES AT GLI GENE MOLECULAR EVOLUTION?

Gunbin K.V.[1,*], Ponomarenko P.M.[2], Ponomarenko M.P.[1], Kolchanov N.A.[1, 2]
[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
[2] Novosibirsk State University, Novosibirsk, Russia.
e-mail: genkvg@bionet.nsc.ru
[*] Corresponding author

Motivation and Aim: We analyze promoters of mammalian key transcription factors genes Gli1, Gli2 and Gli3 responsible for morphogenesis to understand their regulation.

Methods and Algorithms: By Mauve 2.0 tool [1] we aligned multiply the Ensembl contigs containing these genes [2] and, then, calculated TBP/DNA-affinity rates, $-\mathrm{Ln}(K_D)$, upon 70bp-upstream aligned DNA sequences with respect to their known transcription starts in mouse and human using our tool published elsewhere [3], as given in Figure on Gli3 genes.

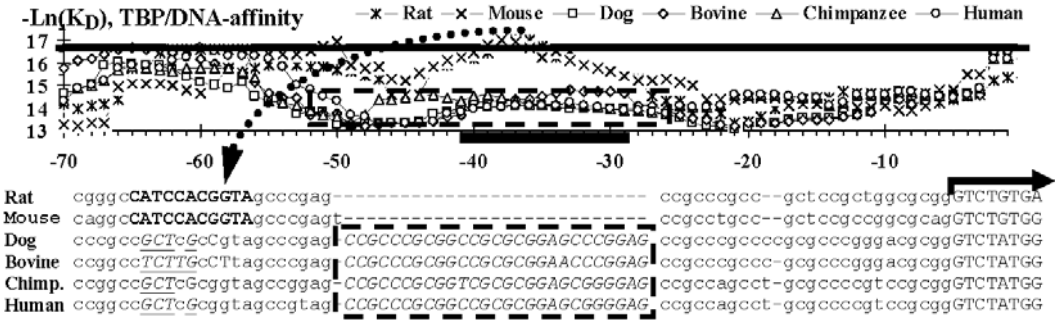


Figure. Gli3 promoters aligned [2] and TBP/DNA-affinity [3] along each of them with upper 95%-confidence cut-off (bold-faced line) that points to (dot-arrow) the rodent TATA-box (bold-faced) hiding at the others mammalians due to substitutions (underlined) and the CpG-island insertions (boxed) in front of transcription starts.

Results and Conclusion: Gli-family promoters in rodents seem to be TATA-containing (dotted arrow) whereas in primates, ungulates, and carnivores they are TATA-less due to the CpG-island insertions (boxed) in-between the transcription start and the TATA-like sequence damaged by substitutions (underlined). As one can see, the rodent-specific TATA-like box (bold-faced) hides at the evolution branch of primates, ungulates, and carnivores Gli-family. This orthologous gene "TATA-containing/less" dichotomy may be associated with commonly accepted dichotomy of the "rapid/slow" maturation of ecological strategies [4]. Indeed, TATA-box proves Gli protein superproduction responsible for the rapid maturation in rodent, R-strategy, whereas K-strategy, slow maturation, of the other mammalians requires making TATA-box "hidden".

Availability: The detailed results are available from the corresponding author upon request.

References:

1. A.C.E. Darling et al. (2004) Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.*, **14**: 1394-1403.
2. T. J. P. Hubbard et al. (2007) Ensembl 2007. *Nucleic Acids Res.* **35**: D610-D617. (Release 48 – Dec 2007)
3. P. Ponomarenko et al. (2008) A stepwise model of TBP/TATA box binding allows for predicting human hereditary diseases by single nucleotide polymorphism. *DAN*, **419**: 88-92.
4. R.H. MacArtur, and E.O. Wilson (1967) The Theory of Island Biogeography, Princeton University Press, Princeton.

MoDELING AND ViSUALISATION OF PATHWAYS USING PETRI NETS

Hariharaputran S., Hofestädt R., Kormeier B., Spangardt S.
Bioinformatics Department, Faculty of Technology, Bielefeld University, Germany
e-mail: hofestae@techfak.uni-bielefeld.de

Motivation: MoViSPP is a new web-based tool to model and visualize biochemical pathways. Today scientific research needs a combination from different data sources. The underlaying infrastructure includes several data sources containing biochemical and metabolic information from Enzyme, MINT, OMIM, PDB, iProClass, GO and KEGG. A relational database management system as backend contains all necessary information. For visualization of the pathway as Petri nets the JUNG (Java Universal Network/Graph Framework) library is used. A web-based graphical user interface is implemented in JavaServer Pages (JSP) that can be accessed with a common web browser.

Implementation: MoViSPP web application consists of a graphical user interface which is useable with a common browser on the client side. Additionally, the application logic and the data warehouse run on server side. The application logic is implemented in platform independent programming language Java.

For the web application MySQL is used as database system on the server side. A JDBC driver for MySQL facilitates the access to the database from the Java classes. JavaServer Pages (JSP) contain Java code that can be included via JSP elements in static HTML code. Hence dependent on the behaviour of the user the static HTML parts can be dynamically be complemented. Whenever a JavaServer Page is called in a web browser a request is send from the client side to the server side. The web container (also called servlet container) on the server side creates, compiles and/or executes a Java servlet. From the server side a response in form of a generated static HTML website is send to the client. MoVisPP uses an Apache Tomcat 6.0.2 servlet container to generate HTML code.

In the web application the user is able to choose different options for the modelling and visualisation of the pathways as Petri nets. There are only a few steps to generate a Petri net. At first the user has to select one of the KEGG pathways. Afterwards the opportunity is given to select a specific organism and to decide if KEGG reactions, the different KEGG relations, interactions from MINT and the data from OMIM should be modelled and visualised. After the previous steps a website with an image of the generated pathway as Petri net (based on the user selection) is displayed. The image map functionality allows the user to get detailed information of all parts of the Petri net.

Application/Results: The semi-automatic construction of Petri net pathways has an added advantage unlike the conventional construction of pathways, using other Petri net tool which is otherwise time consuming and manual intensive. Bringing in the integrated information for the pathways of interest at the same time also providing the possibility to simulate using commercial and non-commercial Petri net based tools provides an edge over already available conventional pathways. If a KEGG pathway is supported with these features it always has an advantage.

For example, gene and protein expression pattern studies of the genes Bcl2 and Tp53 prove their role in cardio vascular diseases. Also, these genes also play a major role in neurodegenerative diseases and in apoptosis. The inputs are generated in the cardio workbench project and are further used as application case. Search for these genes in KEGG database listed the pathways involving these genes.

The figure shows a KEGG pathway based MoViSPP generated map created at run time. The interactive maps are supported with some more information from other in-built and integrated databases such as Enzyme, MINT, OMIM, PDB, iProClass, GO and KEGG which provide more information of the entity involved along with the legend and supported with color indicators as described in the full paper.

Furthermore the maps generated from MoViSPP can be exported and can be used by several other applications that support Petri nets like cell illustrator. The export function provides different possibilities i.e., CSML, PNML and SBML based outputs. Also, the map when imported into cell illustrator allows simulation of the same pathway further making them dynamic in nature allowing user interference.

Conclusion: During the last decade more and more publications showed that the Petri net application seems to be a useful method to solve the fundamental representation and simulation problem of metabolic networks. Petri nets are useful to model concurrent biochemical processes. This allows the natural based representation of biochemical, gene regulation and cell communication processes. Today the most important argument for the usage of Petri nets is that this kind of formalization allows a stepwise extension of the network. That means that we can start with a simple discrete network, which is representing the knowledge of en expert. Furthermore, more details can be included using published papers, databases or information systems like pubmed. In this paper we present a new tool, which allows the semiautomatic construction of metabolic pathways.

WWW-link: http://agbi.techfak.uni-bielefeld.de/movispp/index.jsp

COMPARISON OF ChIP-CHIP SP1 BINDING LOCATION DATA FOR HUMAN CHROMOSOME 21, 22 WITH PWM HITS

Heinzel A.[1, 2,*], Kulakovskiy I.V.[2], Makeev V.J.[2]
[1] Department of Bioinformatics, University of Applied Sciences, Hagenberg, Austria
[2] FGUP GosNIIgenetika, Moscow, Russia
e-mail: mail@aheinzel.at
[*] Corresponding author

Motivation and Aim: Chromatin immunoprecipitation on DNA chips (ChIP-chip) is most commonly used for *in vivo* discovery of transcription factor binding sites (TFBS). However, this technique only locates TFBS within large sequence segments and not yields their precise positions. The spatial resolution of site location may be increased *in silico.* Position weight matrices (PWM)-based prediction is an appropriate method for *in silico* identification of TFBS. Although popular, this technique is suffering from specific tradeoffs and can only give estimated positions. Our aim is to compare ChIP-chip raw signal values and computational predictions. The resulting information may help to develop a combined approach for genome wide identification of TFBS with a high precision.

Methods and Algorithms: In silico predicted Sp1 TFBS with different scores were mapped to *in vivo* observed Sp1 binding segments published in [1]. The number of *in silico* TFBS within and outside of *in vivo* binding segments was evaluated and the overall hit ratio was calculated. The raw intensity of tiling arrays, used for GST and Sp1 detection [1], was quantile-normalized [2], scaled and transformed into signal values. On both datasets, the auto correlation function was applied to calculate the spatial correlation between signal values. Based on the correlation between signal positions the average length of DNA fragments produced by DNA sonication was calculated. In the vicinity of isolated Sp1 TFBS predicted *in silico* the ChIP signal value distribution was evaluated.

Results: The overall number of *in silico* predicted Sp1 TFBS, which are outside of *in vivo* predicted TFBS varies from 3-10% depending on the PWM threshold. The overall hit ratio increases with the increasing threshold, so the enrichment of experimental ChIP-chip segment with "strong" sites is greater than that with "weak" sites. Correlation analysis reveals a very similar length-correlation pattern for both Sp1 and GST datasets having two significant correlation lengths. Thus, it seems that correlation of close signals for Sp1 comes out of length distribution of fragments produced through DNA sonication but not from site positional clustering in the chromosome. This average length can be estimated between 90 - 120 bp with the maximal length around 1200 bp.

Conclusion: Comparison of ChIP-chip and PWM based TFBS prediction results combined with analysis of spatial correlation of signal values can build a basis for a algorithm for better genome-wide TFBS predictions.

References:

1. S. Cawley et al. (2004) Unbiased Mapping of Transcription Factor Binding Sites along Human Chromosomes 21 and 22 Points to Widespread Regulation of Noncoding RNAs, Cell, 116: 499-509.
2. B.M. Bolstad et al. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias, Bioinformatics, 19: 185-193.

DNA – THE PROGRAMMING LANGUAGE OF LIFE?

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

Motivation: During the last decades molecular genetics could identify and sequence different gene functional units (*DNA_Units*). Most of these units are analysed in syntax (sequence and genome) and semantic (metabolic function). This information is growing and represented by different database systems (EMBL – sequences, PDB – structure and semantics etc.), which are available via the internet. Based on this knowledge it is possible to discuss the open question: Can DNA be interpreted as a programming language?

In this paper we will show that the DNA can be interpreted as a programming language in the sense of computer science [1]. Moreover, it is possible to describe the so called "Programming Language of Life" using classical methods of compiler systems.

Method: To show that the *DNA_Units* can be interpreted as a programming language we have to:

a) Specify relevant functional units.
b) Show that the *DNA_Units* represent fundamental mechanisms of a programming language.

In this paper only b) will be discussed, because a) can be done using specific Chomsky-type-2 grammars [2]. The fundamental features of a programming language are:

F1. Data type (at least one is sufficient)
F2. Instruction (standard instructions or by definition)
F3. Control instructions
F4. Punctuation mark

Result/Interpretation: *Let the DNA be the genetic program of a cell. In that case the cytoplasm can be interpreted as the data type representing metabolites. Therefore, we can say that the* **data type** (**metabolite**) *is available (F1).* **Instructions** *are chemical reactions caused by enzymes. Therefore, we can interpret* ***enzymes*** *as instructions which are presented by structured genes (F2). Structure genes are controlling the metabolism indirectly. Regarding specific cells we can see specific genes, which are active during specific time periods. This behaviour shows directly that* **DNA_Units** *can be interpreted as* **control instructions** *(F3). Finally the DNA_Unit which is called* **spacer** *can be interpreted as the* **punctuation mark** *of this system (F4).*

However, based on these ideas this talk/paper will show that the *DNA_Units* represent the features of a programming language and that the *DNA_Units* can be interpreted as a programming language.

References:

1. H.Atlan, M.Koppel (1990) THE CELLULAR COMPUTER DNA: PROGRAM OR DATA, Bull. Of Mathematical Biology, **52**, 335-348.
2. R.Hofestädt (2007) Extended Backus-Systems for the representation and specification of the genome, Journal of Bioinformatics and Computational Biology, 5-2(b): 457-466.

COMPUTER IDENTIFICATION OF GENE TRANSCRIPTION START SITES POSITIONS IN HUMAN, MOUSE, RAT WHOLE GENOME SEQUENCES USING DATA, ANNOTATED IN TRRD

***Ignatieva E.V.*[1, 2]*, *Nechkin S.S.*[1, 2], *Podkolodnyy N.L.*[1, 2, 3]**
[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia;
[2] Novosibirsk State University, Novosibirsk, 630090, Russia;
[3] Institute of Computational Mathematics and Mathematical Geophysics, SB RAS, 630090, Russia
e-mail: eignat@bionet.nsc.ru
* Corresponding author

Motivation and Aim: The TRRD [1] contains information about the transcription regulatory regions in the eukaryotic genes accumulated by manual annotation of scientific articles. TRRD stores data on the experimentally identified transcription start sites (TSSs) of the eukaryotic genes in nucleotide sequences of 100 bp and longer collected in GenBank or EMBL nucleotide sequence database. To integrate TRRD with other genomic databases, it is required to identify the positions of the TSSs in the whole genome sequences. Our aim was to develop a computer technology and to estimate its efficiency using three organisms: *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*

Methods and Algorithms: As a first step, on the basis of data from the TRRD (xml version), we extracted from the entries of the EMBL/Genbank 1705 gene fragments (-1000/+1000 relative to TSS) of three species. Then, using Blat [2], we identified different variants of sequences alignments with fragments of genomic sequences in the neighborhood of the gene starts (GSs) whose positions are given in EntrezGenome. Criteria for choice of the best alignment variant were homology level and total gap length.

Results: Comparison of the TSSs positions we identified in the human genomic sequences relative to GSs stored in EntrezGenome demonstrated that TSSs are not consistently identified at points corresponding to the GSs. In some cases the distance between them may be more than 190 000 bp (the human *MITF*). The disagreement between TSS and GS locations may be due to the presence of alternative TSSs, incomplete data on 5'UTR, sequencing or assembly errors, etc. Search of homologous regions in genomic Builds for three species revealed different characteristics of the data sources. The percentage of EMBL/GenBank sequences for which we succeeded to identify homologous regions in genomic fragments (homology > 95%) was 96.8% for the human, 97.5% for the mouse and 88.7 % for the rat genes. In other situations, the estimates for the alignments and TSS localization on genome sequence should be obtained manually.

Conclusion: The positions of gene TSSs annotated in TRRD for three species were identified. In the course of identification of the TSSs positions on the whole genomes, differences in data sources (*Homo sapiens*, *Mus musculus*, *Rattus norvegicus* NCBI Builds) were determined, and concepts (terms) of TSS and GS were compared. Using the obtained distribution of TSSs relative to gene starts, we developed an accurate procedure for alignment and identification of TSS localizations in the whole genome.

References:

1. N.A. Kolchanov et al., (2002) Transcription Regulatory Regions Database (TRRD): its status in 2002. *Nucl Acids Res*, **30:**312-317.
2. W.J.Kent (2002) BLAT--the BLAST-like alignment tool. *Genome Res,* **12:** 656-664.

COEVOLUTION OF DUPLICATED GENES

***Innan H.* ***
Graduate University for Advanced Studies, Hayama, Japan
e-mail: innan_hideki@soken.ac.jp

Motivation and Aim: Duplicated genes can coevolve by exchanging their DNA fragments through gene conversion. This research aims to detect evidence for coevolution using comparative genomic approaches. To understand the evolutionary background behind coelvolution, theoretical frameworks for analyzing DNA sequences of coevolving duplicated genes are also developed.

Methods and Algorithms: We take advantage of the availability of multiple genomic sequences of baker's yeast and its relatives. The detection of evidence for coevlution involves identification of duplicated genes and simple evolutionary analysis of DNA sequences. We use the diffusion and coalescent approached for developing population genetic theories and analytical and simulation tools.

Results, Conclusion and Availability: Evidence for coevolution was detected in a number of duplicated genes, indicating extensive roles gene conversion in the evolution of duplicated genes in yeast (1). The evolutionary patterns of DNA sequences of duplicated genes are investigated theoretically and by simulations under various conditions (2-5). Simulation tools will be available through the Innan lab web

http://www.sendou.soken.ac.jp/esb/innan/InnanLab/

References:

1. L.Z.Gao, H.Innan (2004) Very low gene duplication rate in the yeast genome. *Science,* **306**: 1367-1370.
2. H.Innan (2003) The coalescent and infinite-site model of a small multigene family., *Genetics*, **178:** 333-444.
3. H.Innan, (2003)A two-locus gene conversion model with selection and its application to the human RHCE and RHD genes. *Proc. Natl. Acad. Sci. USA.*, **100**: 8793-8798.
4. K.M.Teshima, H.Innan, (2004) The effect of gene conversion on the divergence between duplicated genes. *Genetics*, **166**: 1553-1560
5. K.M.Teshima, H.Innan, (2008) Neofunctionalization of duplicated genes under the pressure of gene conversion. *Genetics*, **178**: 1385-1398.

NANOBIOTECHNOLOGY AND ITS APPLICATION TO BIOMEDICINE: TEXT-MINING KNOWLEDGE EXTRACTION AND INTEGRATION

Ivanisenko V.A.*, Demenkov P.S., Yarkova E.E., Ivanisenko N.V., Ivanisenko T.V., Surnina N.Yu., Podkolodniy N.L., Khlebodarova T.M., Ibragimova S.S., Smirnova O.G.
Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
e-mail: salix@bionet.nsc.ru
* Corresponding author

Motivation and Aim: Nanobiotechnology is a dramatically developing sector of science and engineering. Applications of nanobiotechnology in biomedicine includes areas such as drug-delivery systems, labeling, diagnostics, treatment, microfluidics, implants, grafts and biosensors for detecting blood gases, blood sugars, proteins and so on. The rapid increase in the volume of publications on nanobiotechnology issues makes timely the task of development of automated tools for extraction of knowledge from scientific texts. Integration of extracted information about nanomaterial, nanoparticle and nanostructure properties with knowledge in areas of molecular biology, biomedicine and pharmacology may serve as a basis for prediction of new areas for use of the now available nanobiotechnological solutions and search of new promising ones.

Methods and Algorithms: Automated extraction of information about nanobiotechnology from PubMed abstracts was performed using the text-mining methods we developed. For text-mining, we also used the thesaurus we previously developed for the names of proteins, genes, microRNAs, metabolites, biological pathways, diseases, cells, and organisms. Information was integrated by building networks for semantic association coupled literary facts about molecular-genetics regulations, physical interactions, also about associations between nanoobjects, molecular-genetic objects, biological processes and diseases.

Results: Abstracts of scientific publications from PubMed were analyzed. The ANDNanobiotechnology (Associative Network Discovery in the Nanobiotechnology) knowledge base was established. The ANDNanobiotechnology software is equipped with tools for reconstruction and visualization of associative networks. ANDNanobiotechnology provides information about the available nanobiotechnological developments, also ensures analysis of the molecular-genetics systems involved in the function of nanobioconstructs, expression of their pharmacological properties and toxicities in living organisms.

Conclusion: Nanobiotechnology emerges from the physical, chemical, biological, biomedical, and engineering sciences. To develop successfully, nanobiotechnology has to meet the requirement of integration of heterogeneous data from these fields of science. The ANDNanobiotechnology system we developed is pioneering in the area of automated extraction of nanobiotechnological knowledge from texts of scientific publications and their integration with the data on molecular-genetic regulations, molecular interactions and associations observed in living organisms.

Availability: The ANDNanobiotechnology system is available at request to the authors.
Work was supported in part by RFBR: 08-04-91313-IND_a, Government contract FASI №02.514.11.4065, RAS presidium program “Molecular and cellular biology”, “Systems biology: computer and experimental approaches.

PDBSITE DATABASE AND PDBSITESCAN TOOL: TEMPLATE-BASED DOCKING AND RECOGNITION OF FUNCTIONAL SITES IN PROTEIN 3D STRUCTURE

Ivanisenko V.A.*, Ivanisenko T.V., Ivanisenko N.V.
Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
Novosibirsk State University, Novosibirsk, Russia
e-mail: salix@bionet.nsc.ru
* Corresponding author

Motivation and Aim: Recognition of functional sites in proteins is a direct computational approach providing a better understanding of protein biological and biochemical functions. Information about the disposition of the functional sites in protein structure to allows increasing the efficiency of the search of inhibitors of the protein function using methods such as docking and provides acceleration of the development of drugs. An important towards molecular docking is the initial positioning of the ligand in the region of the protein functional site. The aim of the work was a further expansion of the PDBSite database and PDBSiteScan program (Ivanisenko et al, 2004, Ivanisenko et al, 2005), early developed for functional site recognition and site template-based docking.

Methods and Algorithms: The relational PDBSite database was developed using MySQL. The new PDBSiteScan version provides the derivation of protein-ligand complexes from template based docking. To implement docking, we developed an auxiliary database known as the PDBLigand library. The PDBLigand library contains atom coordinates of the low molecular ligands, proteins, DNA and RNA, which bind to the sites from PDBSite. Template based docking is done by transfer of the ligand together with the site-template during the structural alignment of the site-template to protein.

Results: We developed a new version for the PDBSite database that contains 3d templates of various protein functional sites (posttranslational modification, catalytic active, organic and inorganic ligand binding, protein-protein, protein-DNA and protein-RNA interactions) and also a new version of the PDBSiteScan tool ensuring the recognition of functional sites using 3d templates and the creation of molecular protein-ligand complexes relying on template based docking. The generated draft protein-ligand complex can be accepted as an approximation to the further more accurate docking or molecular dynamics analysis. The number of functional and drug binding sites stored in PDBSite was considerably increased, also, the relational version of the PDBSite database was developed.

Availability: The new versions of the PDBSite and PDBSiteScan are now being installed at the IC&G SBRAS server. Work was supported in part by RFBR: 08-04-91313-IND_a, Government contract FASI №02.514.11.4065, RAS presidium program “Molecular and cellular biology”, grant “Systems biology: computer and experimental approaches.

References:

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

MOLECULAR RECOGNITION IN OLIGOMERIC ENZYMES: SUBUNITS AND INHIBITORS INTERACTION

Ivanov A.S. *, Molnar A.A., Mezentsev Yu.V., Ershov P.V., Gnedenko O.V., Archakov A.I.
V.N. Orechovich Institute of Biomedical Chemistry RAMS, Moscow, Russia
e-mail: alexei.ivanov@ibmc.msk.ru
* Corresponding author

Motivation and Aim: Molecular recognition in protein complexes plays a central role in biochemical processes. The interface areas of protein-protein interactions (PPI) have unique structures and represent prospective targets for a new generation of drugs [1-2]. Oligomeric enzymes are the most interesting group of such targets. The aim of this work was investigation of molecular recognition between subunits and PPI inhibitors in oligomeric enzymes.

Methods and Algorithms: We have chosen two oligomeric enzymes as test molecular objects most convenient for such research - HIV-1 protease (HIVp) (homo-dimer) and bacterial L-asparaginase (homo-tetramer) [3]. We have used some computer methods: molecular modeling, computational alanine scanning, molecular dynamics simulation and energy minimization. All calculations were done using Sybyl 6.9.1 (Tripos Inc.) and Amber 7 software running on SGI Origin200 server. For experimental measurement of molecular interactions SPR-biosensor Biacore-3000 was used.

Results: Computer analysis of subunits contact areas in HIVp dimer and homotetramer of bacterial L-asparaginase was done using virtual «alanine scanning». Several amino acid residues ("hot spots") which bring maximal contribution into the interaction energy have been found. In case of L-asparaginase The basic attention was given to the interface between dimers AC and BD and between monomers in these dimers. It was shown, that in each subunit there are 13 residues which play a key role in interaction between dimers AC and BD. Computer and biosensoric analysis of intersubunits interactions were carried out

Specificity of PPI in oligomeric enzymes was tested on the set of close relative bacterial L-asparaginases from: *E. coli, Erw. chrysanthemi, Erw. carotovora and H. pylori* (26695 and J99). The estimation of possible formation of chimeric hetero-complexes by subunits from different asparaginases was done using computer simulation and SPR-biosensor. The highest degree of molecular recognition between subunits has been shown.

Thermodynamics of subunits interaction in both enzymes, as well as interaction inhibitors of PPI [4] were also studied.

Acknowledgments: This work was supported in part by Russian Foundation for Basic Research (grant 07-04-00575).

References:

1. A.V. Veselovsky et al. (2002) Protein-protein interactions: mechanisms and modification by drugs. J. Mol. Recognit., 15: 405-422.
2. A.S. Ivanov et al. (2005) Bioinformatics Platform Development: From Gene to Lead Compound, Methods Mol. Biol., 316: 389-432.
3. Yu.V. Mezentsev et al. (2007) Oligomerization of L-asparaginase from Erwinia carotovora. Biochemistry (Moscow) Suppl. Ser. B: Biomedical Chemistry, 1(1): 58-67.
4. A.S. Ivanov et al. (2007) Protein-protein interactions as new targets for drug design: virtual and experimental approaches. J. Bioinform. Comput. Biol., 5(2b): 579-592.

SELECTION OF NEW TARGET PROTEINS FOR DRUG DESIGN IN GENOME OF *MYCOBACTERIUM TUBERCULOSIS*

Ivanov A.S.*, Molnar A.A., Veselovsky A.V., Skvortsov V.S., Archakov A.I.
V.N. Orechovich Institute of Biomedical Chemistry RAMS, Moscow, Russia
e-mail: alexei.ivanov@ibmc.msk.ru
* Corresponding author

Motivation and Aim: The big number of currently used antimicrobial drugs do not meet medico-biological requirements - not safety application and efficiency reduction in connection with occurrence of drug-resistant microbial strains. Way out of this situation is creation of the new antimicrobic drugs with action, distinct from already known. It means, that they must influence on new molecular targets and therefore the searching of new targets for antimicrobic drugs is actual. The aim of this work was the predicting of new potential targets for antitubercular agents.

Methods and Algorithms: Recently, we have used bioinformatics approaches for predicting new potential targets for antitubercular agents [1]. Genomes of *M. tuberculosis H37Rv*, *M. tuberculosis CDC1551*, *Mycobacterium leprae*, sequences of all known human proteins and proteins from PDB were utilized for these purposes. Experimental measurement of protein-ligand interactions was carried out with optical biosensor Biacore-3000.

Results: S election of target proteins encoded by genome of *M. tuberculosis* H37Rv was done based on the following criteria: 1) no homologues in human; 2) identical both in *M. tuberculosis H37Rv* and *CDC1551* strains; 3) similar to proteins from *M. leprae TN* (>67% of homology); 4) homologues in PDB (>40% of identity). Preliminary hit list of targets (13 proteins) was analyzed in detail and target prioritization, based on probable protein functions, was carried out. As a result, only 8 proteins were selected as potential targets (rpoA, rpsD, rpsE, prsH, kdtB, ruvA, and kdtB). Later, two of them were also found by other investigators as potential targets for wide-spectrum antibacterial agents: phosphopantetheine adenylyltransferase (PPAT) (kdtB) and DNA helicase (ruvA).

Later we selected PPAT for the next step of research - finding new inhibitors of PPAT using virtual and experimental screening of compounds from chemical databases of commercial available samples. High purity PPAT and high resolution (1.6 A) 3D structure of PPAT were obtained from our collaborators - Institute of Bioorganic Chemistry RAS and Institute of Crystallography RAS, correspondingly.

Based on SPR-biosensor technology we created test system for experimental screening of potential lead compounds. This work will be continued.

Acknowledgments: This work was supported in part by Russian Federal Space Agency (in frame of ground preparation of space research).

References:

1. A.V. Dubanov, A.S. Ivanov, A.I. Archakov (2001) Computer searching of new targets for antimicrobial drugs based on comparative analysis of genomes. Vopr. Med. Khim. 47, 353-367. (in Russian).
2. A.S. Ivanov, A.V. Veselovsky, A.V. Dubanov, V.S. Skvortsov (2005) Bioinformatics Platform Development: From Gene to Lead Compound, Methods Mol. Biol., 316: 389-432.

LINK OF EXON AND INTRON LENGTHS IN ANIMAL GENES

Ivashchenko A.T.*, Atambaeva S.A., Khailenko V.A., Achsheulov A.S.
Kazakh National University named after al-Farabi, Almaty, Kazakhstan
e-mail: a_ivashchenko@mail.ru
* Corresponding author

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

Methods and Algorithms: The genes of chromosome 1 *Gallus gallus*, *Canis familiaris* and *Equus caballus*, chromosome 5 *Danio rerio* and *Monodephis domestica* were arranged in samplings with 1, 2, 3, 4, 5, 6-9, 10-14, 15 and more introns in a gene. In each sampling of genes mean of exon (l_{ex}), intron (l_{in}) and gene length (L_{gn}), sum of exon lengths (L_{ex}) and intron number in a gene (N_{in}) was calculated.

Results: Exon and intron lengths decreased in some times while augmenting intron number in genes of fish and hen. These variations of exon and intron lengths are interdependent and they are described by linear regressions: $l_{ex} = 7.3\ l_{in} + 1298$ with a coefficient of correlation (r) equal 0.840 (*D. rerio*) and $l_{ex} = 12.2\ l_{in} + 1458$ with r equal 0.898 (*G. gallus*). The exon length essentially changes in genes of the horse, dog and opossum when increasing intron number. The sum of exon lengths in genes of the horse, dog, opossum, hen and fish increased accordingly in 4.3, 4.2, 3.4, 4.7 and 4.2 times when increasing intron number in genes. The augmentation of exon lengths sum changes proportionally to increasing of intron number in genes. This link is described by linear equations with the high value of r:

E. caballus: $N_{in} = 0.0082\ L_{ex} - 4.29$ (r =0.990), $N_{in} = 0.00032\ L_{gn} - 1.18$ (r =0.994);
C. familiaris: $N_{in} = 0.0081\ L_{ex} - 4.53$ (r =0.993), $N_{in} = 0.00026\ L_{gn} + 0.08$ (r =0.990);
M. domestica: $N_{in} = 0.0088\ L_{ex} - 6.18$ (r =0.988), $N_{in} = 0.00017\ L_{gn} - 0.40$ (r =0.995);
G. gallus: $N_{in} = 0.0079\ L_{ex} - 3.40$ (r =0.996), $N_{in} = 0.00033\ L_{gn} - 1.09$ (r =0.997);
D. rerio: $N_{in} = 0.0082\ L_{ex} - 3.69$ (r =0.988), $N_{in} = 0.00050\ L_{gn} - 1.63$ (r =0.996).

The gene length of the horse, dog, opossum, hen and fish when increasing intron number changes accordingly in 20, 19, 16, 12 and 11 times. The gene length augmented with the increase of intron number in genes and this link is described by a linear regression, which testifies to the above mentioned.

Conclusion: The obtained results demonstrate that the eukaryotic intron-containing genes have an identical principle of exon-intron structure. When increasing protein-coding nucleotude sequience there are inserts introns in a gene. The dissection of this sequience in exons by means of introns takes place while observing of quantitative rations between exon and intron lengths.

References:

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

COMPUTATIONAL MODELLING OF GENE REGULATION IN THE CNIDARIANS *NEMATOSTELLA VECTENSIS* AN *ACROPORA MILLEPORA*

Kaandorp J.A.*, Nanfack Y.F., Postma M.
Section Computational Science, Faculty of Science, University of Amsterdam, Amsterdam,
e-mail: jaapk@science.uva.nl
* Corresponding author

Motivation and Aim: Within the metazoans, sponges and cnidarians represent the phyla with the simplest body plan and a relatively simple regulatory network controlling the development. This makes these organisms an excellent case study for understanding morphogenesis and the physical translation of the genetic information into a growth form, using a combination of biomechanical models of growth and form and a model of the spatial and temporal expression of developmental genes.

Methods and Algorithms: We use a set of coupled partial differential equations to model spatio-temporal gene expression patterns. We use a cell boundary model to model cell morphology, cell adhesion and cell movement in early development of the cnidarians *Nematostella vectensis* and *Acropora millepora.* An optimalisation method [1] based on evolutionary algorithms is applied to infer gene networks and model parameters from gene expression data obtained from in situ hybridizations.

Results: A method for modelling and simulation of gene regulation in combination with cell movement and adhesion.

Conclusion: We can simulate several aspects of gene regulation and cell movement in early development in *Nematostella vectensis* and *Acropora millepora*

References:

1. Y. Fomekong Nanfack, J.A. kaandorp, J.G Blom (2007) Efficient parameter estimation for spatio-temporal models of pattern formation: case study Drosophila melanogaster, *Bioinformaticsl,* **23:** 3336-3363.

EXON-INTRON STRUCTURE OF *D. DISCOIDEUM*, *T. PARVA* AND *P. FALCIPARUM* GENES

Kabdullina A.A., Ivashchenko A.T.*
Kazakh National University named after al-Farabi, Almaty, Kazakhstan
e-mail: a_ivashchenko@mail.ru
* Corresponding author

Motivation and Aim: Exon-intron structure of lower eukaryotes genes has not investigated insufficiently. It is necessary to bring out if there is some link between the changes of exon and intron length from intron number that of higher eukaryotes genes [1]. Studying the properties of intron-containing genes of lower eukaryotes will promote to clarification of biological role of introns.

Methods and Algorithms: The genes of completely sequencing genomes of *Dictyostelium discoideum*, *Plasmodium falciparum* and *Theileria parva* were arranged in samplings with 1, 2, 3, 4, 5, 6-9, 10-14, 15 and more intron in a gene. In each sampling of genes was calculated mean of exon (l_{ex}), intron (l_{in}) and genes length (L_{gn}), sum of exon lengths (L_{ex}) and intron number in a gene (N_{in}). The amount of introns and exons with the length at the intervals 1-20, 21-40, 41-60 n and so on up to 400 n as well as a length more than 400 n has been analyzed.

Results: In *P. falciparum, D. discoideum* and *T. parva* genes the l_{ex} decreased accordingly in 5.8, 3.9 and 4.0 times at the augmentation N_{in} from 1 to 17. When increasing intron number the l_{in} decreased. The link between changes of the l_{ex} and the l_{in} has been expressed by the following linear equations: $l_{ex} = 0.110\, l_{in} + 117$ (*P. falciparum*), $l_{ex} = 0.077\, l_{in} + 97$ (*D. discoideum*) and $l_{ex} = 0.110\, l_{in} + 65$ (*T. parva*).

While increasing intron number the sum of exon lengths of *P. falciparum, D. discoideum* and *T. parva* genes enlarged in 1.6, 2.3 and 2.3 times according to linear equations: $N_{in} = 0.0104\, L_{ex} - 18.8$ ($r = 0.89$), $N_{in} = 0.0038\, L_{ex} - 4.1$ ($r = 0.78$) and $N_{in} = 0.0095\, L_{ex} - 9.1$ ($r = 0.97$).

In *P. falciparum* genes a close link between the length of genes and intron number in them has been observed expressed by the equation: $N_{in} = 0.0048\, L_{gn} - 9.9$ ($r = 0.99$). The close link between the L_{ex} and N_{in} caused close link between L_{gn} and N_{in} which in *D. discoideum* and *T. parva* genes is described by the equations: $N_{in} = 0.0022\, L_{gn} - 2.5$ ($r = 0.88$) and $N_{in} = 0.0049\, L_{gn} - 4.3$ ($r = 0.99$).

Changes of exon lengths at the augmentation of N_{in} in *P. falciparum* and genes descends in such a way that a lobe of exons with the length more than 400 n decreases in 10 times and a lobe of exons with a length at an interval 60-140 n increases in 3 times. The value l_{in} decreases too due to reducing a lobe of introns with the length of more than 400 n from 10.8 to 0.3% and the augmentation of a lobe introns at an interval of 100-140 n from 24 to 60%. The change of l_{ex} at the augmentation of N_{in} in *T. parva* genes descends so that a lobe of exons with a length of more than 400 n decreases in 9 times and a lobe of exons with the length at an interval of 80-160 n increases in 3 times. The l_{in} decreases too due to reducing a lobe of introns with the length of more than 400 n from 4.5 to 0.5% and the augmentation a lobe of introns at an interval 40-80 n from 50 to 68%.

Conclusion: The obtained results demonstrate that intron-containing genes of lower and higher eukaryotic organisms have an identical rule of exon-intron structure.

References:

1. S.Atambaeva, V.Khailenko, A.Ivashchenko (2008) Changes of introns and exons length in genes of arabidopsis, rice, nematode and human, *Mol. Biol.* (Russ), **42:** 1-10.

SELECTIVITY OF ALLELE SPECIFIC HYBRIDIZATION OF DNA PROBES

Kabilov M.R.*, Pyshnyi D.V.
Institute of Chemical Biology and Fundamental Medicine, SB RAS, Novosibirsk, Russia
e-mail: kabilov@niboch.nsc.ru
* Corresponding author

Motivation and Aim: The method of molecular hybridization of oligonucleotide probes with nucleic acids used for revealing specific sequences has found a wide utility in molecular biology. The technologies allow not only the revealing of specific nucleotide sequences but also discrimination of minimal perturbations in the nucleic acid structure, such as, for example, single mismatches, which leads to destabilization of duplex formation. The providing of the highly selective interaction between nucleic acids and oligonucleotides or their analogs and derivatives is an important task for physico-chemical biology and, in particular, for DNA diagnostics. To date, however, there is no generally accepted parameter for the quantitative estimation of the hybridization selectivity.

Methods and Algorithms: A simple model of allele specific hybridization has been viewed when an oligonucleotide probe interacts with two templates, one of which is completely complementary and the other one contains one nucleotide substitution, which leads to the decreased efficiency of complex formation. Thermodynamic characteristics (dH^o and dS^o) for complementary and mismatched complexes were calculated for the standard conditions (1M NaCl) using unified nearest neighbor parameters. The selectivity function, that is the ratio of the association extents of complementary and noncomplementary complexes, was used for the description of the hybridization selectivity. This function corresponds to the ratio of specific and nonspecific signals upon hybridization.

Results: The notions of the probe selectivity for the specific hybridization conditions and, in general, the limit probe selectivity, should be clearly distinguished. The limit probe selectivity depends solely on temperature and ddH^o and ddS^o whereas the selectivity function is determined, in addition, by the length and concentration of the probe. The influence of all this parameters on selectivity hybridization has been considered. The temperature corresponding to the maximal selectivity is slightly higher than melting temperature of the complementary complex. Unexpectedly, the short probes were shown to have the lower maximal selectivity than the long ones.

Conclusion: In this work we considered the theoretical aspects of the selectivity of hybridization of oligonucleotide probes with nucleic acids. The proposed functions of the probes selectivity allow evaluating the efficiency of discrimination of any perturbation in an analyzed NA duplex. These perturbations can be natural (mismatches) or caused by some artificial modifications, e.g., by introduction of a non-canonical base. The obtained results can be the basis for the software for the choice of the optimal probe structure, which has the maximal selectivity under the given conditions, and for the strategy of the analysis of new probe types selectivity.

This work was supported by MCB Program of RAS (10.6), integration grant of SB RAS (55, 73), RFBR grant (06-04-49263), and SS (3689.2008.4).

DISCOVERY OF PROPERTY-CONSERVED FUNCTIONAL ELEMENTS IN HUMAN

Kalybaeva Y.M.[1]*, Deyneko I.V.[1,2], Blöcker H.[1], Kauer G.[3]
[1] Department of Genome Analysis, HZI, Braunschweig, Germany;
[2] ICG, Novosibirsk, Russia;
[3] University of Applied Sciences, Emden, Germany. e-mail: yka04@helmholtz-hzi.de
[*] Corresponding author

Motivation and Aim: Conservation across the genome sequences of several organisms and/or within one genome has helped to discover many novel functional sequence elements. However, the one weak point of this classical approach is that it relies on conservation of nucleotide letters only, while nature is "function-oriented".

Methods and Algorithms: To overcome this weak point of the classical approach we presented a novel, signal-theoretical technology (1). To evaluate conservation of DNA properties rather than just primary sequence, we applied our novel tool (2) and carried out phylogenetic comparisons of promoter regions of genes from human, mouse, rat and dog. Starting from the compiled dataset of promoter sequences (–2000bp…–1bp), we searched for motifs (10-15bp), which are conserved in the four genomes and statistically over-represented in this dataset. In contrast to other studies (3), motifs were regarded as similar if they showed similar property profiles, which were calculated using the DNA conformational parameter "roll" and the DNA physical parameter "melting enthalpy". Although the matching sequences may slightly vary in letters, they nevertheless represent identical properties in real nature.

Results: We then compiled a top list of motifs (104 in total), many of which were similar to known binding sites for different transcription factors. Normalized by the total length of all found motifs, the frequency of some predicted binding sites (PWM taken from TRANSFAC) in the top list is ~10^2 times higher than compared to observed frequencies in promoter regions of human genes (-2Kbp…-1bp from TSS) or exons. In comparison to a similar work, based on nucleotide letter statistics (3), our top-scoring list is similar, but not identical and contains in addition, for example, binding motifs for HMG-I(Y), SRY and Pbx1a transcription factors.

Conclusion: We suppose that for these three motifs (and for many others identified by our approach, but not by letter conservation) the more important characteristic is conformation of the DNA double strand, rather than simple nucleotide conservation. In general, our approach offers a broader view of similarity (with 38 DNA parameters at hand) and may help to pull-out many novel functional elements for which no classical consensus can be found.

References:

1. Kauer G. and Blöcker H. (2003) Applying signal theory to the analysis of bio-molecules. *Bioinformatics*, **19**, 2016-2021.
2. Deyneko IV. *et al*. (2006) FeatureScan: revealing property-dependent similarity of nucleotide sequences, *Nucleic Acids Res*., **34**, W591-595.
3. Xie X. *et al*., (2006) Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature*, **17**, 338-345.

GENE MIGRATION AND WORD FLOW IN INDONESIA

Karafet T.M.[1, 2*], Lansing J.S.[2], Hammer M.F. [2]
[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia;
[2] University of Arizona, Tucson, USA
e-mail: tkarafet@email.arizona.edu
[*] Corresponding author

Motivation and Aim: We examined large-scale patterns of single nucleotide polymorphism (SNP) variation on the Y chromosome of Indonesian people 1) to investigate the origins and relationships of Indonesian populations; 2) to study the history of the colonization of Indonesia, Melanesia and Australia; 3) to relate processes of genetic and linguistic change.

Methods: Eighty five Y-chromosome SNPs and 14 STRs (Short Tandem Repeats) were typed by standard methods. The software package ARLEQUIN 3.0 was used to calculate population parameters and to perform Mantel test. Quantitative measurement of distance between all pairs of languages was made using ALINE distance. The isolation-with-migration (IM) model was used to estimate two daughter populations split from a constant-sized ancestral population, but with continuing migration between the two demes (http://lifesci.rutgers.edu/~heylab/HeylabSoftware.htm#IM).

Results: For a long period of time most of Indonesian territory was a dry land. Islands from Western Indonesia formed a direct extension of the Asian mainland – Sundaland, while Australia and New Guinea formed a single prehistoric continent, called Sahulland. The area between Sundaland and Sahullalnd, named Wallacea, has never formed a continuous land bridge between Asia and Australia. To understand the genetic affinities among indigenous Indonesians, their Y-chromosome data were compared with populations from Southeast Asia and Pacific region. We found that Indonesian islands differed to a great extent by their Y-haplogroup composition. As a result their Fst value was 0.42 – one of the highest Fst's for any regions in the world. Western Indonesian islands shared intensively Y-haplogroups with Mainland Asia whilst Eastern Indonesians were similar to Papuans. Three dominant haplogroups in Eastern Indonesia yielded dates older than 30,000 years ago, these estimates were consistent with Pleistocene heritage of these lineages. We examined language change in a contact zone where incursive agricultural communities interacted, and subsequently merged, with aboriginal foraging societies on the eastern Indonesian island of Sumba and found a strong correlation between language and genes.

Conclusions: Significant distinction in genetic structure between Western and Eastern Indonesia reflects long time of separation between Sundaland and Wallacea. Considerable genetic differences between Australians and New Guineans might be due very old and separate initial populations from Wallacea. We demonstrate gene flow to Western Indonesia from the northern populations prior to Austronesian expansion. Historical patterns of social interaction reconstructed from genetic, linguistic and climatic data largely explain community-level language evolution on one of Indonesian island.

Acknowledgments: This research was supported by the National Science Foundation to JSL, MFH and TMK.

SEARCH FOR PLANT HOMOLOGUES OF ANIMAL STRUCTURAL MICROTUBULE-ASSOCIATED PROTEINS IN THE *ARABIDOPSIS THALIANA* GENOME

Karpov P.A. [*]***, Blume Y.B.***
Institute of Cell Biology and Genetic Engineering NAS of Ukraine,
acad. Zobolotnogo str., 148, Kiev, 03680, Ukraine
e-mail: karpov.p.a@gmail.com
[*] Corresponding author

Motivation and Aim: Microtubule-associated proteins (MAPs) play an important role in the regulation of microtubule (MT) structure and functions. They stabilize MTs and support their interactions with other cell components. Most of the plant MAPs discovered so far have homologues in eukaryotes, but not all animal or fungal MAPs are present in plants. Although because of MAPs role in modulating microtubule functions, the full complement of these important sequences has not yet been described.

Methods and Algorithms: The sequences of known structural MAPs were obtained from Swiss-Prot/TrEMBL, GenBank and DDBJ. The FASTA-sequences were sorted accordingly to their types and families (MAP1, MAP2_tau, MAP4). The information on domain architecture and functional regions was obtained from the Swiss-Prot and was analysed with a help of SMART. Alignment was performed in Clustal X (1.8). Motifs (not less than six amino acid residues) were considered conservative if 100%-identity was observed only with typical animal MAPs under scanning in the SIB BLAST. Identification of *Arabidopsis thaliana* homologues was accomplished with the NCBI tBLASTn using complete protein sequences of domains or tubulin binding repeats. The sequence handling, assembly, translation and search of the *A. thaliana* genome were carried out with the help of different Lynnon BioSoft tools. Localization of consensus regions, corresponding to the candidate genes were carried out using the NCBI Map Viewer and data represented in Tair, TIGR, KEGG, ExpPASy and GenBank.

Results:. The gene-identification strategy, utilizing site-based and comparative methods, allowed us to identify amino acid motifs that are conserved within each MAP family, and designed corresponding nucleotide motifs in IUPAC code, accommodating potential codon polymorphism by applying the universal symbols representing all possible substitutions. Scanning the complete sequences of *A. thaliana* chromosomes, we identified candidate coding sequences for plant homologues of animal structural MAPs. Furthermore, we identified 200 loci (i.e., 178 of MAP1 and 22 of MAP2 and tau) that are of interest as potential chromosomal regions for plant structural MAPs. Interestingly, consensus regions for MAP4 were not identified in *A. thaliana*. Chromosomal locations for three previously known plant homologues of MAP1 (i.e., AtEB1a, AtEB1b and AtEB1c) were also identified in the *A. thaliana* genome.

Conclusion and Availability: Based on the results of investigations, we inclines to assumption, that if the new homologues of MAP1 will be detected within the genome regions designated, their sequences will soon differ considerably from the sequences of typical animal MAPs. Nevertheless, slight sequence resemblance of the not exclude existence of functional homology. In the case of MAP2_tau and MAP4 we assume the possibility of existence of some proteins, functionally similar to MAP2. At the same time, if plant homologues do exist, the sequences similarity will be very low. In the case of MAP4, our results confirm the absence of MAP4 in plants. In any case, if plant homologues of animal MAP2_tau and MAP4 do exist, their similarity will be very weak. So, there divergence probably took place at a very early stage in their evolution.

MALE/FEMALE DISCRIMINATION: USAGE OF PROBE-LEVEL AFFYMETRIX EXPRESSION DATA

Karyagina A.S.[1]*, Ershova A.S.[1, 2], Nurtdinov R.N.[2], Vasiliev M.O.[3], Merkov A.B.[4], Lossev I.S.[5]

[1] N.F. Gamaleya Institute of Epidemiology and Microbiology, Moscow, Russia; [2] M.V. Lomonosov Moscow State University, Moscow, Russia; [3] Moscow Institute of Physics and Technology, Moscow, Russia; [4] Institute for System Analysis, Moscow, Russia; [5]Parascript LLC, Boulder, USA

e-mail: akaryagina@gmail.com

* Corresponding author

Motivation and Aim: Standard Affymetrix technology evaluates gene expression by measuring the intensity of mRNA hybridization with olygonucleotide probes; the probes are grouped into probe sets. The probes within a probe set may work very differently, but standard technique uses only integral characteristics of probe sets. We propose to consider each probe individually that provides us with larger amount of data and may lead to better results. The approach is implemented to the model of male/female discrimination using expressions of genes in sex chromosomes.

Methods and Algorithms: CEL files obtained using Affymetrix U133 Plus 2.0 Array for different cancer tissues were extracted from GEO repository (GSE2109 entry), each file has the male/female label. Our aim was to reveal the probes and genes allowing to separate sexes and to use the probes for the description of the samples. First, each probe of each sample was assigned with a numerical value; we used the PM-MM differences that give better results than PM values. Second, for each probe we have found the threshold giving the minimal number of wrongly attributed files, i.e. the male files belonging to the female part of values with respect to the threshold and *vice versa.* Third, the probes having the above error number below 5% of the total number of files were selected as the *informative* probes. Finally, we describe each file with a Boolean vector, each component of the vector corresponds to a selected probe (0 if it is below the threshold and 1 otherwise). To map Affymetrix probes on human genome we have constructed and used an Affy_AltSplice DataBase (http://affymetrix.bioinf.fbb.msu.ru/).

Results: For all probe sets of X and Y chromosomes only 18 probes from X chromosome and 27 probes from Y chromosome were selected as the informative ones. Those probes belong to two probe sets from X chromosome (both from the XIST gene) and five probe sets from Y chromosome (corresponding respectively to the genes RPS4Y1, EIF1AY, DDX3Y, JARID1D, CYorf15B). Each of above 7 probe sets contains both informative and non-informative probes. Clusterization of Boolean vectors shows several clusters that can be interpreted as (a) normal males (551) and females (1251), (b) males (19) and females (19) with wrong sex labels, (c) males with high expression for all chosen probes of Y-chromosome and high XIST expression (10), (d) males with several unexpressed probes for genes of Y chromosomes (15), (e) females with XIST probes unexpressed (22). The minor subgroups are subject of biological interpretation. E.g. male samples with pronounced X-chromosome inactivation (group (c)), are often observed in testis tumors, as well as in patients with Klinefelter's syndrome as a result of X-chromosome amplification.

Conclusion: Different probes within a probe set have different levels of reliability. The suggested approach to reveal and use informative probes provides biologically reasonable results.

MGSmodeller – A COMPUTER SYSTEM FOR RECONSTRUCTION, CALCULATION AND ANALYSIS MATHEMATICAL MODELS OF MOLECULAR GENETIC SYSTEM

Kazantsev F.V.[1, 3]*, Akberdin I.R.[1], Bezmaternykh K.D.[1, 2], Lashin S.A.[1], Podkolodnaya N.N.[1], Likhoshvai V.A.[1, 2]

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

[2] Novosibirsk State University, Novosibirsk, Russia

[3] High College of Informatics of the Novosibirsk State University, Novosibirsk, Russia

e-mail: kazfdr@bionet.nsc.ru

[*] Corresponding author

Motivation and Aim: Development of multifunctional analytical computer system that presents resources for modelling and analysis of the functional patterns in molecular genetic systems of eu- and prokaryote cells is one of the current important problems in system biology. In this study the MGSmodeller computer system that intended for reconstruction, calculation and analysis mathematical models of molecular genetic system is presented.

Methods and Algorithms: MGSmodeller as a computer system contains the next set of modules: “model constructor/editor”, “model calculation”, “inverse problem”, “optimal control”. Multifunctionality of the system is based on the facilities of each module. Such module as “model constructor/editor” allows to create and to edit models using original standard SiBML for specification mathematical models of molecular genetic networks. SiBML based on generalized chemical kinetic method [1] and taking into account the structure of their genetic and compartmental levels.. The system also includes: (i) tools for model reconstruction with arbitrary structure of the molecular genetic systems taking into account positional relationship and orientation genes in genomes, polyallel genes, matrix principles of the passing such fundamental processes as replication, transcription and translation, and milticompartmental structure of studying systems, (ii) tools for dynamic calculation, solving of inverse problems and problems of optimal control. MGSmodeller is java-application and is supplied with attributes of the user friendly interfaces. User interface of the MGSmodeller allows to hierarchically represent data, edit them and to clearly represent the calculation results and to perform analysis of molecular genetic models.

Results: The original format of models description SiBML, algorithms and the MGSmodeller system are developed for generation, reconstruction, edition and calculation of molecular genetic networks models. This system was used for reconstruction the auxin metabolism mathematical model [2]. All information relatively MGSmodeller and about authors is available on the web-site http://www.bionet.nsc.ru/labs/modelgroup/indexEng.html

References:

1. V.A. Likhoshvai et al. (2001). Generalized chemokinetic method for gene network simulation, *Mol. Biol.*, **35:**1072-1079.
2. I.R. Akberdin et al, Mathematical model of auxin metabolism in shoots of arabidopsis thaliana L, *this volume*

MGSgenerator – THE TOOL FOR AUTOMATICAL GENERATION OF MOLECULAR GENETIC SYSTEM MATHEMATICAL MODELS ON BASIS OF GENE NETWORKS STRUCTURE

Kazantsev F.V.[1, 3]*, Akberdin I.R.[1], Bezmaternykh K.D.[1, 2], Likhoshvai V.A.[1, 2]
[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
[2] Novosibirsk State University, Novosibirsk, Russia
[3] High College of Informatics of the Novosibirsk State University, Novosibirsk, Russia
e-mail: kazfdr@bionet.nsc.ru
* Corresponding author

Motivation and Aim: One of the topical and interesting challenge in the system biology is creating automatic tool for generation of the mathematical models on basis of reconstructed gene networks with further analysis of this complex models and visualization of model results. Since reach this aim we have developed special software that allow to convert information about gene network structure to the mathematical model. We called this system as MGSgenerator.

Methods and Algorithms: MGSgenerator is based on Eclipse (www.eclipse.org) plug-in architecture. The developed converter support following types of plug-ins. First is the data source modules and second is the data export modules. Thus, this tool can be extended for conversion from different systems such as KEGG data base and for generation models in different modeling systems such as *Mathematica*. The reaction rate is described on basis of Hill functions, that allow to represent different types of the molecular-genetic regulation. It's computer system could be used for generation mathematical models only for correct described subset of patterns that could be representing in gene networks. Anyway it might be networks where are complex processes, which couldn't be described by pre-set rules. In this case user can manually set the necessary reaction rate equation.

Results: This software tool MGSgenerator could be and is used to generate math models of gene network patterns subset. Today we have few plug-in modules that allow us to convert format XML of the GeneNet computer system [1] into SiBML format of the MGSmodeller [2]. MGSmodeller is the complex software which allows to reconstruct models and, of course, to analyze and calculate mathematical models. The gene network that we used for MGSgenerator testing didn't allow us to describe for less then 20% of molecular-genetic processes by the mathematical models, but after gene network annotations modification we converted all processes in this gene network in the SiBML. All information relatively MGSgenerator and about authors is available on the web-site http://www.bionet.nsc.ru/labs/modelgroup/indexEng.html.

References:

1. Ananko E.A., Podkolodny N.L., Stepanenko I.L., Podkolodnaya O.A., Rasskazov D.A., Miginsky D.S., Likhoshvai V.A., Ratushny A.V., Podkolodnaya N.N., Kolchanov N.A. (2005) GeneNet in 2005. Nucleic Acids Res., 33: 425-427.
2. Kazantsev F.V. et al. MGSmodeller – a computer system for reconstruction, calculation and analysis mathematical models of molecular genetic system, *this volume*

PREDICTION OF THE REGULATORY MECHANISMS OF *ESCHERICHIA COLI YFIA* GENE EXPRESSION

Khlebodarova T.M.*[1], Likhoshvai V.A.[1, 2], Oshchepkov D.Y.[1], Kachko A.V.[1, 2], Tikunova N.V.[1, 2]

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

[2] Novosibirsk State University, Novosibirsk, Russia

e-mail: tamara@bionet.nsc.ru

*Corresponding author

Motivation and Aim: Progress in sequencing of the genomes from different organisms raised the problem of rapid functional decoding of their genetic programs. One possible approach could be based on computer analysis of potential regulatory regions of the genes, mathematical modeling of their functions and experimental verification of the putative mechanisms. This approach allows mass analysis of the sequences. We used this approach to reconstruct the structure of the potential promoter of *E. coli yfiA* gene. Previously, we have constructed a genosensor based on the potential *yfiA* promoter and the *gfp* reporter gene. The sensor cells *E.coli*/pYfi-gfp responded to various external toxic stimuli, including oxidative stress [1]. However, the mechanism underlying the response remains unknown.

Methods: Mathematical models were developed by the method of generalized Hill functions [2]. Recognition of the potential binding sites for the transcription factors (TF) was done using the SITECON method [3]. Protein-DNA binding was analyzed by EMSA (electrophoretic mobility shift assay).

Results: Modeling of the response on oxidative stress in sensor cells *E.coli*/pYfi-gfp demonstrated that the coincidence with the experimental data is the best in the model implying a complicated effect presumably through several TFs. With regard to this fact, we searched for TF potential binding sites (TFBS) in the *yfiA* promoter region. Totally, 52 alignments of TFBS were analyzed. The high recognition level estimated by type II error (false positive) was demonstrated for 19 sites belonging to 16 TFs. Among them, several binding sites for TFs (MarA, IscR, MetJ, PurR, and SoxS), directly or indirectly involved in the response of this gene to oxidative stress, were found. In addition, and some potential TFBSs regulating the metabolism of sugars, amino acids, nucleic and fatty acids were discovered. Moreover, the presence of the *Crp* TFBS, which is a global regulator of the genes responsible for sugars catabolism, was proved by EMSA.

Conclusion: Detection of the great number of potential sites for TFs, both for oxidative agents and global regulators of metabolic processes, supports our assumption that the *E.coli yfiA* gene is a sensor of metabolism disorders in a cell. So, the genosensor derived from its promoter could be a polyfunctional sensor.

References:

1. N.V. Tikunova et al. (2007) A computational-experimental approach to designing a polyfunctional genosensor derived from the *Escherichia coli* gene *yfiA* promoter. *Dokl. Biochem. Biophys.*, **417**: 357-361.
2. V. A. Likhoshvai, A.V. Ratushny (2007) Generalized hill function method for modeling molecular processes. *J. Bioinform. Comput. Biol.*, **5:** 593-610
3. D.Y. Oshchepkov et al. (2004) SITECON: a tool for detecting conservative conformational and physicochemical properties in transcription factor binding site alignments and for site recognition. *Nucleic Acids Res.* **32:** W208-W212.

DISCOVERY OF THE TRANSCRIPTION FACTOR BINDING SITES IN THE ALIGNED AND UNALIGNED DNA SEQUENCES

Khomicheva I.V.[1, 2], Vityaev E.E.[2], Shipilov T.I.[2]
[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia,
[2] Sobolev Institute of Mathematics, SB RAS, Novosibirsk, Russia
e-mail: khomicheva@bionet.nsc.ru
[*] Corresponding author

Motivation and Aim: Most of the available tools for transcription factor binding site (TFBS) prediction are based on simplifying assumption of independent contribution of each nucleotide position to the binding affinity. The assumption is not universally valid. The task of developing the method overcoming this limitation is the challenging one.

Methods and Algorithms: 'ExpertDiscovery' system discovers the specific nucleotide patterns of arbitrary length that could be characterized as (1) statistically significant; (2) hierarchically complicating; (3) allowing gaps [1].

Results: We analyzed the DNA targets for several protein families (CEBP, HNF4, SF1 an so on). In the case when the *apriori* alignment of TFBSs is known, 'ExpertDiscovery' outperforms position weight matrix (PWM) at any threshold cutoff (Fig. 1, a). When the *apriori* alignment is not known, 'ExpertDiscovery' performs much better than PWM at any threshold cutoff (Fig. 1, b).

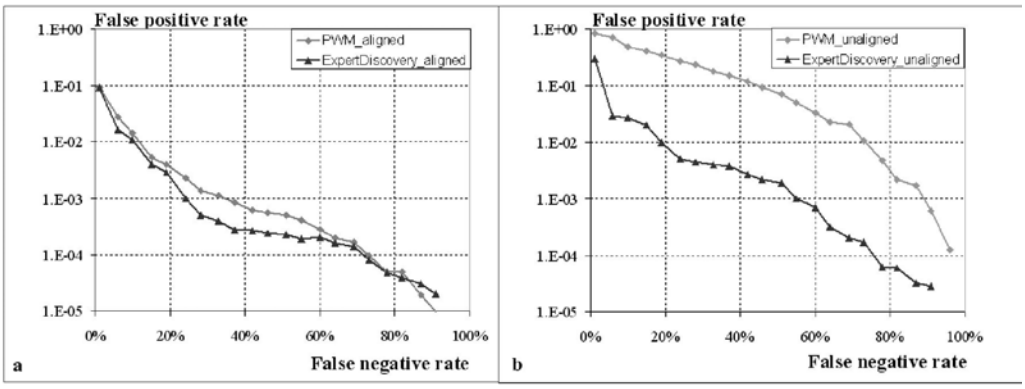


Figure 1. False positive versus false negative rates calculated for 'ExpertDiscovery' system and PWM using bootstrap procedure to recognize CEBP binding sites in two cases: a. aligned data, b. unaligned data.

Conclusion: TFBS sequences possess the complex degenerate structure, they vary in length, position, redundancy, orientation in the DNA chain. Some factors seem to show position dependencies whereas others do not. 'ExpertDiscovery' finds the common building patterns of unaligned TFBSs and allows to predict the new members of family factors.

Availability: http://www.math.nsc.ru/AP/ScientificDiscovery/pages/projects.html

References:

I. Khomicheva et al. (2008) ExpertDiscovery application for the hierarchical analysis of the eukaryotic transcription regulatory regions based on the DNA codes of transcription, *Intelligent Data Analysis*, in press.

PREDICTION OF PROTEIN INTERACTIONS USING HOMOLOGOUS INTERFACES

Kirys T.V.*[1, 2], Tuzikov A.V.[1], Voytekhovsky D.K.[1], Grushetsky Y.E.[1]
[1] United Institute of Informatics Problems BAS, Minsk, Belarus
[2] Belarusian State University, Minsk, Belarus
e-mail: kirys@newman.bas-net.by
[*] Corresponding author

Motivation and Aim: The importance of protein in all living systems is immense. At the protein level most biological mechanisms are based on shape-complementarity, so that proteins present particular concavities and convexities that allow them to bind to each other and form complex structures. In general, proteins perform their functions by forming complexes. Therefore the knowledge of interactions between proteins is essential for understanding the molecular mechanisms of biological systems and drug design. The aim of our research is the prediction of protein interactions using homologous interfaces. We address two problems. First, given an interface database and a protein database the task is to predict possible protein interactions. Second, given a target protein, an interface database and a protein database the task is to find possible interacting partners.

Methods and Algorithms: The assumption behind the homology-based approaches is that interaction information can be extrapolated from one complex structure to homologs of the interacting proteins. Define an interface as a pair of interacting binding sites (patch) of two protein. Each patch consists of all residues of a protein that are located within a distance of $6\,\mathring{A}$ from the other protein. From such an interface definition it is clear that in a patch there are continuous segments of protein polypeptide chain. We use that observation for searching similar patches in proteins. We also assume that if a protein spatial feature is similar to a patch spatial feature then the corresponding distance matrices are highly correlated.
The proposed algorithm consists of the following steps:

1. select continuous segments in the interface;
2. find similar continuous segments in proteins using dynamic programming on distance matrix;
3. superpose proteins by the correspondence found;
4. check modeled interaction (steric clash, similarity score);

Taking into account residue types makes prediction more accurate.

Results: This algorithm was implemented in C++. Our algorithm is quite fast, depending on the protein and interface sizes it takes several seconds to find the solution, what is important in screening task.

Conclusion: We have proposed a novel homology-based algorithm for prediction of protein interactions, which employs dynamic programming on distance matrix. Finding particularly spatially similar continuous segments reduces searching time drastically in comparison with hashing algorithms and adequately reflects the nature of protein interactions.

Availability: The software is available on request from the authors.

MODELLING OF REGULATORY NETWORKS TO INDENTIFY PROMISING DRUG TARGETS FOR BREAST CANCER THERAPY

Koborova O.N.[1], Filimonov D.A.[1], Zakharov A.V.[1], Lagunin A.A.[1], Kel A.[2], Kolpakov F.[3], Sharipov R.[3,4], Kondrachin Y.[3], Poroikov V.V.[1]

[1] Institute of Biomedical Chemistry of Rus. Acad. Med. Sci., Moscow, Russia;
[2] BIOBASE GmbH, Germany;
[3] Institute of Systems Biology, Novosibirsk, Russia;
[4] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia;
e-mail: okoborova@gmail.com

Motivation and aim: Cell cycle regulation abnormalities lead to serious diseases. One of the most wide spread is breast cancer. There are some drugs acting on different targets, but despite of that only in US in 2007, approximately 40,460 women are expected to die from breast cancer (American Cancer Society). Thus, finding of new drug targets for cancer therapy, is important task.

Methods and algorithms: We propose an algorithm for anticancer drug target identification. The algorithm models cell cycle regulation as logic networks. Input data is the initial states of nodes (proteins and/or genes) in a primary moment of time. Output data is a number of node states in different time moments – trajectories. For analysis of proteins, influence on which can change trajectories in a desired way (i.e., apoptosis stimulation of cancer cells without influence on normal cells or stop of cell cycle progression), the node states can be fixed according to the type of proposed effect (inhibition of the respective proteins).

Results: The method was applied to the case of breast cancer using molecular network consisting of 234 nodes: 361 proteins from TRANSPATH database (http://www.biobase.de), and expression data for breast cancer, consisting of up and down regulated genes list from Cyclonet database (http://cyclonet.biouml.org). The network consists of two functional blocks: cell cycle block and stress associated pathway block. Two groups of promising targets were identified: first - a number of targets, which inhibition changes trajectory into interruption of cancer cells division, and second - a number of targets, which inhibition changes trajectory into apoptosis of breast cancer cells.

Conclusion: Proposed algorithm evaluated the anti-tumor targets and its combination taking into account the microarray data for breast cancer. Preliminary results demonstrate the applicability of the method to identify new targets.

Availability: According to Net2Drug Consortium agreement.

Acknowledgement: The work was supported by European Commission project Net2Drug No. 037590 (FP6-2005-LIFESCIHEALTH-7).

X-CHROMOSOME INACTIVATION, MOBILE ELEMENTS AND ncRNA GENES

Kolesnikov N.N.*, Elisafenko E.A.
Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
e-mail: kolesnikov@bionet.nsc.ru
* Corresponding author

Motivation and aim: X-chromosome inactivation is the epigenetic phenomenon which occurs in female mammals, whereby one X chromosome is silenced in every cell. X-inactivation, together with genomic imprinting, has provided an important model system for studying epigenetic regulation, i.e. regulation occurring apart from the DNA sequence itself. Indeed, the transcriptional inactivation of the one X chromosome is a complex process, controlled by the inactivation center, denoted as XIC. The XIC is a complex genetic locus comprising a number of genes producing non-coding nuclear RNAs. Within the XIC the non-coding (nc) RNA gene, Xist, is the key player. Xist is virtually unique in that it is the only known gene where the ncRNA product, and not a protein, epigenetically regulates a whole chromosome. Previous studies showed that two of Xist's exons originated from a protein coding gene. In this study we have conducted an independent analysis of the origin and evolution of the inactivation center and *Xist* gene in eutherians using a set of bioinformatics approaches, the impact of mobile elements on the creation of the ncRNA genes.

Results: In this study we showed that key XIC gene *Xist***,** which display fragmentary homology to the chicken protein coding gene *Lnx3*, emerged de novo ~ 180 Mya by integration of mobile elements and colonization of simple tandem repeats. The *Xist* gene promoter region and four of ten exons found in eutherian retain homology to exons of *Lnx3* gene. The rest six *Xist* exons including simple tamdem repeats detectable in their structure have similarity to different transposable elements. Integration of mobile elements into *Xist* accompanies the overall evolution of the gene and presumably continues in contemporary eutherian species. Additionally we noted that combination of remnants of protein-coding sequences and mobile elements is not unique for the *Xist* gene and found in the other XIC genes producing non-coding nuclear RNA**.** As a result additional data demonstrating how the XIC originated from a region containing protein-coding genes have been obtained. In particular, we have demonstrated that the genes *Enox*(*Jpx*) and *Ftx* of the XIC, contain exons homologous to those in cognate protein-coding genes *Uspl* and *Wave4(Wasf3)*, moreover, different types of dispersed repeats give birth to the exons of the genes in different mammalian species.

Conclusion: Thus, the eutherian X-inactivation center has originated from a region containing protein-coding genes that has undergone considerable structural rearrangements, invasion of mobile elements and processed pseudogenes, and pseudogenization of protein-coding genes.

The key XIC gene, *Xist*, displays a fragmentary homology to the chicken protein-coding gene *Lnx3* and, most likely, emerged *de novo* during formation of the eutherian X chromosome ~180 Mya. Consequently, XIST can be classified as an intermediate pseudogene and this gene evolution has contributed to the creation of a new ncRNA gene. Our data, adds to the story that Xist originally stemmed from a protein coding gene by showing that a transposable elements inserted within the gene contributed to the creation of a ncRNA gene. In fact, we see that other regions of the XIC have evolved similarly by transposable element insertion, mutating active genes into non coding pseudogenes. Our data constitute important findings into the origin of the unique Xist gene and others ncRNA genes of the XIC.

CYCLONET - AN INTEGRATED DATABASE ON CELL CYCLE REGULATION AND CARCINOGENESIS

Kolpakov F.A.[1, 2*], Poroikov V.V.[3], Sharipov R.N.[4, 1, 2], Milanesi L.[5], Kel A.E.[6]
[1] Institute of Systems Biology, Novosibirsk, Russia; [2] Design Technological Institute of Digital Techniques, SB RAS, Novosibirsk, Russia; [3] Institute of Biomedical Chemistry RAMS, Moscow, Russia; [4] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia; [5]Institute of Biomedical Technologies, CNR, Segrate (MI), Italy; [6]BIOBASE GmbH, Wolfenbuettel, Germany
* e-mail: fedor@biouml.org

Motivation and Aim: The aim of the Cyclonet database [1] is to develop an integrative approach that will help researchers to understand the cell cycle regulation and carcinogenesis process through modelling and simulation of gene regulatory networks.

Methods and Algorithms: Modern software technologies were used for Cyclonet development. BioUML technology (http://www.biouml.org) was applied for the formal description of structure and functioning of complex biological systems and processes on different logical levels represented in diagrams, as well as for visual modelling and simulation of eukaryotic cell cycle using built-in Java engine. BeanExplorer Enterprise Edition (http://www.beanexplorer.com) was used for Cyclonet integration with free international biological databases and development of web interface for user access via Internet. The BMOND database (http://bmond.biouml.org) was used as repository of diagrams, description diagram elements (genes, proteins, substances, concepts, reactions and semantic relationships), mathematical models and results of simulation.

Results: Cyclonet integrates data of genomics, proteomics, chemoinformatics, and systems biology to use them in design of more effective drugs:

• genomics – 354 links to available microarray experiments were collected and categorized. This section also contains complete list of human genes derived from five microarray experiments and 30 lists of up- and down-regulated genes for different subtypes of breast cancer revealed by meta-analysis of these experiments, as well as three lists of genes characterized by monotonical expression in cell cycle;

• proteomics – information about 3436 proteins, their complexes and interactions;

• chemoinformatics – information about 55 key targets for anticancer treatments, 62 anticancer pharmacological activities, 422 corresponding activities, and 4335 chemical sustances with chemical formulas and described physiological activities. This information is used by PASS system (http://www.ibmc.msk.ru/PASS) to predict new ligands with anticancer activities;

• systems biology – more than 150 diagrams describing cell cycle regulation and related systems and 32 mathematical models of cell cycle regulation annotated from literature.

Cyclonet is also integrated with BMOND (http://bmond.biouml.org) – the specialized database on cell cycle regulation and microarray data. Both databases are used for storage and exchange of data collected in the frameworks of international project FP6 "Net2Drug".

Availability: Cyclonet is available at http://cyclonet.biouml.org.

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

References:

1. F.A. Kolpakov et al. (2007) CYCLONET - an integrated database on cell cycle regulation and carcinogenesis, *Nucleic Acids Res,* **35** (Database issue)**:** D550-556.

IDENTIFICATION OF TRANSCRIPTION FACTORS MEDIATING ENTRY INTO CELLULAR QUIESCENCE

***Kondrakhin Yu.V.*[1, 2], *Sharipov R.N.*[1, 3, 2], *Filipenko M.L.*[4], *Boyarskikh U.A.*[4]**
[1] Institute of Systems Biology, Novosibirsk, Russia; [2] Design Technological Institute of Digital Techniques, SB RAS, Novosibirsk, Russia; [3] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia; [4] Institute of Chemical Biology and Fundamental Medicine, SB RAS, Novosibirsk, Russia
*e-mail: yvkondrat@mail.ru

Motivation and Aim: Regulation of quiescence may be critical to degenerative diseases and cancer. Serum deprivation of fibroblasts was studied in [1] as a model of entry into cellular quiescence. It remains a challenge to identify transcriptional factors that regulate over-expression of genes after serum deprivation. In the frameworks of this work we aimed at the two main tasks: 1) identification of genes over-expressed after serum deprivation; 2) analysis of promoter regions of identified genes with the purpose to reveal the transcription factors that significantly regulate them in conditions of serum deprivation.

Methods and Algorithms: For identification of genes over-expressed after serum deprivation we used the statistical method based on optimization of parameters of hyper-geometrical distribution. For recognition of potential binding sites of transcription factors we used the weight matrix method described in [2]. Initial frequency matrices were extracted from the TRANSFAC (http://www.biobase-international.com/pages/index.php?id=transfac) and JASPAR databases (http://jaspar.genereg.net). Statistical simulation was applied to decrease the number of false positive results.

Results: Analyzing the microarray data we selected two groups of genes. The 1st group contained the genes significantly over-expressed after serum deprivation. The 2nd (control group) contained the genes with not changed expression. Promoter [-1000bp, +1000bp]-regions of all selected genes were extracted from the Ensembl database (http://www.ensembl.org). Promoter regions of genes from both groups were analyzed by the matrix method for recognition of transcription factor binding sites. Comparison of the occurrence frequencies of recognized binding sites allowed to reveal a set of transcription factors (for instance, KLF4, PPARG and FOXF2), which responded to serum deprivation. Application of chi-squared independence test allowed us to reveal additional set of transcription factors, whose binding sites significantly co-occured in promoters of the genes from the 1st group. The most striking examples were IRF2, c-Ets-1 (p54), PU1 and STAT6.

Acknowledgements: SB RAS Integrational Grant n13 “Stem cells for future biotechnology”.

References:

1. H.Liu et al. (2007) A transcriptional program mediating entry into cellular quiescence, *PloS Genetics,* **3:e91:** 1-13.
2. E.A.Ananko et al. (2007) Recognition of interferon-inducible sites, promoters, and enhancers, *BMC Bioinformatics*, **8:56:** 1-14.

COMPARISON OF BREAST CANCER AND COMMON CANCER on THE BASE OF META-ANALYSIS OF MICROARRAY DATA

***Kondrakhin Yu.V.*[1, 2], *Sharipov R.N.*[1, 3, 2], *Kolpakov F.A.*[1, 2*]**
[1] Institute of Systems Biology, Novosibirsk, Russia; [2] Design Technological Institute of Digital Techniques, SB RAS, Novosibirsk, Russia; [3] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia;
*e-mail: fedor@biouml.org

Motivation and Aim: Cancer is a highly heterogeneous disease. It remains a challenge to identify common transcriptional features of cancer as well as particular features of different cancer subtypes. Currently, a lot of gene expression experiments were intended for studying various types and subtypes of cancer. Public available microarray data sets were accumulated in such databases as the Stanford Microarray Database (http://genome-www5.stanford.edu) and ArrayExpress (www.ebi.ac.uk/arrayexpress). Meta-analysis of microarray data on gene expression in normal and cancer cells from independent experiments could allow to reveal reliably differentially expressed genes. In the frameworks of this work we aimed at several main tasks: 1) meta-analysis of five data sets to identify significantly the genes expressed differentially in breast cancer; 2) comparison of identified genes with genes revealed by meta-analysis in [1] as differentially expressed in common cancer; 3) comparative analysis of promoter regions of identified genes with the purpose to reveal the transcription factors that significantly participate in neoplastic progression.

Methods and Algorithms: For identification of differentially expressed genes we have developed the statistical method based on optimization of parameters of hyper-geometrical distribution. For recognition of potential binding sites of transcription factors we used the weight matrix method described in [2]. Initial frequency matrices were extracted from the TRANSFAC database (http://www.biobase-international.com/pages/index.php?id=transfac).

Results: Our meta-analysis has identified significantly (p-value<0.001) that 431 genes were differentially expressed in breast cancer tissues. 187 genes dysregulated in nearly all cancerous tissues were identified in [1]. About 70% of these genes were also differentially expressed (at least, insignificantly) in breast cancer. Promoter [-1000bp, +1000bp]-regions of genes up-regulated in breast and common cancer were analyzed by the matrix method for recognition of transcription factor binding sites. Promoter regions of selected genes were extracted from Ensembl database (http://www.ensembl.org). We discriminated a set of transcription factors (for instance, PAX6, COUP, TBX5 and ABI4), which were characterized as specific for breast cancer only. We also identified the set of transcription factors (for instance, NF-Y, YY1, E2F and ICSBP) as specific for common cancer.

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

References:

1. Y.Lu et al. (2007) Common human cancer genes discovered by integrated gene-expression analysis, *PloS ONE,* **11:** 1-13.
2. E.A.Ananko et al. (2007) Recognition of interferon-inducible sites, promoters, and enhancers, *BMC Bioinformatics*, **8:56:** 1-14.

ANALYSIS OF THE CORRELATED MUTATIONS IN THE HOMOLOGOUS PROTEINS OF THE FPG/NEI FAMILY

***Koptelov S.S.*[1*]*, Afonnikov D.A.*[1, 2]*, Zharkov D.O.*[1, 3]
[1] Novosibirsk State University, Novosibirsk, Russia
[2] SB RAS Institute of Cytology and Genetics, Novosibirsk, Russia
[3] SB RAS Institute of Chemical Biology and Fundamental Medicine, Novosibirsk, Russia
e-mail: skoptelov@ya.ru
[*] Corresponding author

Motivation: Organisms repair their damaged DNA in several ways, one of which, base excision repair (BER), removes most of nucleobase lesions. The first step of BER is performed by DNA glycosylases, such as bacterial enzymes Fpg and Nei, which catalyze excision of many oxidatively damaged bases from DNA.

To investigate which structural features are important for the functions of enzymes, phylogenetic methods can be used. One of such methods is the detection of correlated mutations. If amino acid (AA) residues interact physically or functionally, they are likely to change in an interdependent manner. Given a set of proteins of a homologous family, the co-evolution can be detected by analyzing the physicochemical properties of residues for correlations. A covariation matrix can be built using the values of property in different proteins and positions; then Pearson's linear correlation coefficients with significance levels can be estimated.

Methods and Algorithms: In this work, a correlated mutation analysis was performed on the 124 AA sequences of the Fpg and Nei proteins from different organisms. Alignment and phylogenetic tree were built (MUSCLE [1], Phylip [2]), and the data was analyzed by the CRASP program [3] for correlated mutations. Results of the CRASP output were further accompanied by the 3D structure analysis to find the pairs of most functional and structural importance.

Results: Significant correlations were detected in residue pairs for isoelectric point (pI) and hydrophobicity. It is shown that several pairs detected by pI correlation analysis evolve by compensatory mutations and form salt bridges in the 3D structure of Fpg. Some of these pairs are located in functionally important sites and may contribute to protein's glycosylase function, e.g. the pair R54/E131 binds N- and C-domains of the enzyme and can significantly influence the structural flexibility of the DNA-binding groove. It is suggested that anticompensatory mutations in these sites will decrease or ablate the glycosylase activity of the enzyme, while the compensatory mutations will not affect it significantly. The hypothesis about structural and functional relationships between these residues is subject for further verification by site-directed mutagenesis.

The work was partly supported by the RAS program "Biosphere origin and evolution", RFBR grant 08-04-00596 and SB RAS integration project 49.

References:

1. Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput, *Nucleic Acids Research*, **32:** 1792-1797.
2. Felsenstein, J. (1989) PHYLIP - Phylogeny Inference Package. *Cladistics*, **5**: 164-166.
3. Afonnikov D.A, Kolchanov N.A. (2004) CRASP: a program for analysis of coordinated substitutions in multiple alignments of protein sequences,*Nucleic Acids Res*,**32**: W64–W68.

CENTRALITY ANALYSIS OF GENE REGULATORY NETWORKS

***Koschützki D.*[1, 2]*, Schreiber F.*[1*]
[1]Leibniz Institute of Plant Genetics and Crop Plant Research, Gatersleben, Germany
[2]Furtwangen University of Applied Sciences, Furtwangen, Germany
[*]Corresponding author, email: schreibe@ipk-gatersleben.de

Motivation and Aim: The increasing quality, size and complexity of gene regulatory networks enforces the application and further development of network analysis methods for their investigation, and various approaches have been developed or employed from other fields of sciences to investigate these complex networks. The ranking of network elements, often called *centrality analysis*, is one of these methods. A centrality is a function which considers the network structure and assigns every vertex of the network a numeric value in a way that more important vertices (such as global regulators in a gene regulatory network) get higher centrality values.

Results: Here we discuss and compare different centrality measures which can be used to analyze gene regulatory networks. Some of these centralities have been already studied in biological sciences; others are transferred, for example, from social network analysis. This study includes the centralities *degree*, *closeness*, *radiality*, *integration*, *shortest path betweeness*, *Katz status index*, and *PageRank*. We analyze these centralities within the gene regulatory network of *Escherichia coli* based on the data of transcriptional regulatory interactions of genes from RegulonDB, Version 5.5 and compare the ranking of the genes based on the centralities with global regulators proposed by Martínez-Antonio & Collado-Vides (2003). Some centralities are able to identify more than 50% of the global regulators within the top 2% of the ranked genes. However, they also result in very different rankings with low pair-wise Kendall's correlation coefficients for several pairs of centralities. We also show that the consideration of biological knowledge can improve the results of centrality analysis (the identification of global regulators). We discuss and extend a complex class of centrality measures, *motif-based* centralities, which is based on network motifs, small recurring sub-networks within a given network which often relate to specific biological functions.

UBIQUITOMIX DATABASE: A NEW RESOURCE ON UBIQUITIN SYSTEM

Kovalyov V.A.[3]*, Gainullin M.R.[1, 3], Eremin E.V.[2], Garcia A.[1]
[1] Nizhny Novgorod State Medical Academy, [2] Institute of Applied Physics RAS, [3] Nizhny Novgorod State University, Nizhny Novgorod, Russia
e-mail: vladlen-85@list.ru
* Corresponding author

Motivation and Aim: Ubiquitylation is a process of great importance for many vital cell functions (proteolysis, signal transduction, control of gene expression, DNA repair, etc). Ubiquitin system consists of complex of enzymes wich catalyse covalent attachment of ubiquitin to a target protein (E1 - ubiquitin activating enzymes, E2 – ubiquitin conjugating enzymes, E3 – ubiquitin protein ligases), particular proteins serving as substrates for ubiquitin modification, deubiquitylation enzymes (DUB) and proteins containing ubiquitin binding domains (UBP). Biological significance of ubiquitin dependent regulation makes it attractive object of research, using methods of systems biology. However among specialized biological Internet-resources (SwissProt/TrEMBL, GO, KEGG, Reactome, etc.) there are no resources correctly and completely describing ubiquitin system. Therefore Internet-resource is required for accumulation and ordering of knowledge on ubiquitin system. The aim of our work was to create specialized object-oriented database, collecting data on all members of ubiquitylation system and their interactions in various organisms.

Methods and Algorithms: For development of Ubiquitomix database BioUml platform has been chosen. It is open source software, being used for formalization of biological systems, as well as for their visualization (http://www.biouml.org).

Results: Main principles of formalization and graphic display of all compounds of ubiquitin system and their interactions have been developed. Proteins were formalized according to their properties. Basing on protein structure we had divided them in two different kinds of entities: simple entities (monomeric or homo-oligomeric proteins) and modular entities (hetero-oligomeric proteins and protein complexes). According to their function, all components of ubiquitin system have been divided into: 1) enzymes of conjugation (i.e. attachment of ubiquitin molecule to target protein); 2) deubiquitylating (the reversal of conjugation) enzymes; 3) proteins which recognize a specific ubiquitin signal; 4) ubiquitin target proteins. For modular entities, components have been divided into 5 types: possessing catalytic activity, carrying out ubiquitin binding function, substrate binding proteins, adaptor proteins and effector proteins. Function and structure of proteins are displayed by form and color of pictogram. Each protein is characterized by name and synonyms, gene name, MW, length, posttranslational modifications, function in ubiquitin system. Additionally, ubiquitylated proteins are characterized by type of ubiquitin chain, site of ubiquitylation and ubiquitylating machinery. If ubiquitylation mechanism of target protein is described completely, respective reaction is displayed on own pathway diagram. For partially characterized ubiquitin binding processes semantic diagrams are chosen. Information on 383 human and 508 yeast proteins has been collected in Ubiquitomix database and their interactions have been analyzed.

Conclusions: General information on ubiquitin system members and their interactions has been collected in developed database. We expect that further development of Ubiquitomix database will be of common interest for different research groups involved in studies of ubiquitin system.

THE NOVEL INTEGRATIVE APPROACH FOR INVESTIGATION OF HIGHLY REPETITIVE SEQUENCES

Krasikova A.V.*, Gaginskaya E.R.
Biological Research Institute of Saint-Petersburg Sate University, Saint-Petersburg, Russia
e-mail: chromas@paloma.spbu.ru
* Corresponding author

Motivation and Aim: Chicken (*Gallus gallus domesticus*) genome is now one of the best investigated bird genomes. Available resources include a consensus genetic map, BAC libraries representing several genome-equivalents organized in physical maps (http://www.bioinformatics.nl/gbrowse/cgi-bin/gbrowse/) and the second draft of the chicken genome sequence. Notwithstanding the progress of the research, the chicken genome sequence still contains numerous gaps (accessible at http://www.ensembl.org, http://ncbi.nih.gov). For a number of microchromosomes no sequence information is available. Among certain difficulties - cloning, sequencing and chromosome assembling of tandem repeat clusters, which constitute centromere, telomere and other heterochromatic regions of chromosomes. Consequently, the DNA sequences of centromeric regions are usually unknown and are represented by gaps in the current chicken chromosome sequence assembly (symbolized by 1.5 Mb gaps for macrochromosomes and 0.5 Mb gaps for microchromosomes). At the same time so-called virtual chromosome (chromosome unknown, ChrUn) containing sequences with unknown chromosome position is highly enriched with tandem repeats.

The aim of the present work was to introduce a novel 'cytogenomic' approach which would provide essential information about the regions absent from the sequence maps of chromosomes including centromeres, telomeres and other heterochromatic regions enriched with tandem repeats.

Results and Conclusions: The integrative approach, which includes computer based analysis and physical mapping at of tandemly repetitive sequences high resolution level, was developed [1, 2]. Examination of sequences bordering heterochromatic regions in the current chromosome sequence assembly allowed us to identify several novel tandem repeat families in the chicken genome including PO28 and PO446 repeats. We also screened the chicken genome databases (ChrUn) to analyze sequences bordering previously described tandem repeats CNM and PO41; 107 contigs containing CNM and/or PO41 repeat clusters were found. Detailed analysis of these DNA fragments with RepeatMasker (http://www.repeatmasker.org), RepBase Update (http://www.girinst.org) and BLAST resources allowed us to predict that satellite transcription could be initiated at long terminal sequences (LTR) of LTR-retrotransposones.

This work was supported by the Russian Foundation for Basic Research (grant 08-04-01328). We used the equipment of the Core Facility "CHROMAS" (Biological Research Institute of Saint-Petersburg Sate University).

References:

1. Krasikova A., Deryusheva S., Galkina S., Kurganova A., Evteev A., Gaginskaya E. (2006) On the positions of centromeres in chicken lampbrush chromosomes, *Chromosome Res,* **14**: 777-789.
2. Deryusheva S., Krasikova A., Kulikova T., Gaginskaya E. (2007) Tandem 41 bp repeats in chicken and Japanese quail genomes: FISH mapping and transcription analysis on lampbrush chromosomes, *Chromosoma,* **116**: 519-530.

SABIO-RK: INTEGRATING REACTION KINETICS DATA FOR SYSTEMS BIOLOGY

Krebs O.*, Wittig U., Kania R., Weidemann A., Mir S., Golebiewski M., Rojas I.
EML Research gGmbH, Heidelberg, Germany
e-mail: olga.krebs@eml-r.villa-bosch.de
* Corresponding author

Motivation: The simulation of biochemical reaction networks requires information about the kinetics of the biochemical reactions participating in the network, such as the kinetic laws describing the dynamics of the reactions with their respective parameters determined under certain experimental conditions. These data are widely scattered through various publications and described in many different formats. Thus, it is a prerequisite for integrating such data into simulation networks to define standards for reporting and exchanging the data obtained, both from experimentalist to modelers and for the feedback from modelers to experimentalists.

Results: SABIO-RK (System for the Analysis of Biochemical Pathways - Reaction Kinetics) is a database designed to store and offer access to information about biochemical reactions and their kinetics in a comprehensive and standardised manner. It stores the fundamental information about biochemical pathways, like reactions and their participants (enzymes, compounds, modifiers). It also offers support for the storage of information about proteins, protein complexes and genes, all this linked to organism (including strains) and to biochemical reactions (in the case of enzymes).

Availability: SABIO-RK can be accessed through a web-based user interface (http://sabio.villa-bosch.de/SABIORK/) that allows the search for biochemical reactions and their kinetics by specifying characteristics of the reactions of interest (such as reactants, enzymes or pathways) as well as the kinetic data searched (e.g. from a particular tissue, determined under certain experimental conditions or only certain parameter types). All reactions matching the search criteria are presented, allowing the view of further details about reactions, catalyzing enzymes and the kinetic data upon selection. Data about biochemical reactions and their kinetic parameters with their respective rate equations can be exported in SBML format (Systems Biology Mark-Up Language, thus allowing the import into simulation and modeling programs supporting SBML

OCCURRENCE OF RECOGNITION SITES OF RESTRICTION-MODIFICATION SYSTEMS IN BACTERIOPHAGE GENOMES

Krivozubov M.S.[1], Ershova A.S.[1], Karyagina A.S.[2, 3], Spirin S.A.[4], Alexeevski A.V.[4]*
[1] Bioengineering and Bioinformatics Faculty, Moscow State University, Moscow, Russia
[2]Gamaleya Institute of Epidemiology and Microbiology, Moscow, Russia
[3] Institute of Agricultural Biotechnology, Moscow, Russia
[4] Belozersky Institute of Physical-Chemical Biology, Moscow State University, Moscow, Russia
e-mail: aba@belozersky.msu.ru
* Corresponding author

Many bacteria hold restriction-modification (RM) systems preventing host cells from invasions of foreign DNA by cleavage the latter at specific sites. Typically, type II RM-system includes two proteins: the restriction endonuclease cleaves foreign DNA at specific short sequences, while the methyltransferase protects host DNA from restriction by methylation of such sites. Bacteriophages possess several strategies to prevent DNA cleavage. For instance, in their genome they can avoid sites recognized by the host RM systems. The degree of conformity of the real number of RM-site occurrences in a phage genome to the expected (assuming their random distribution) number of such sites, points out the presence of an RM-system against this sequences in the phage host cell and allows to propose other phage strategies against RM-systems.

In the present work, we have computed the number of occurrences of 259 known type II RM-sites in 366 bacteriophage genomes and determined the expected occurrence number by using Markov model estimation. We have found 2535 pairs (phage genome, RM-site) with the underrepresented site, i.e., such that the number of sites is significantly less than statistically expected.

A comparison of underrepresented phage sites and host RM-systems has confirmed aforesaid supposition, but also has revealed some cases when an appropriate RM-system in the host cell is unknown. We suggest two possible explanations of this situation: either the host really has a necessary RM-system, but it is still not annotated, or the phage can have one or more additional hosts with a necessary RM-system besides its annotated hosts. The second variant can be confirmed by taxonomic or ecological proximity of annotated and possible additional phage hosts. In particular, we have determined 297 cases, in which the bacteria genus of the annotated phage host includes also a bacterium possessing an appropriate RM system.

Also we have divided all 2545 genome-site pairs into two groups. The first group contains pairs with zero or very few sites per genome. This group represents cases of evident phage survival against the host RM-system. The second group contains pairs with the number of sites that is less than statistically expected but, nevertheless, can reach several tens or hundreds per genome. The latter group is quite large, while the mathematical model predicts an exponential fall of chances to survive depending on the absolute number of sites in a genome.

Proceed from the contradiction, we propose that the phages from the second group either recently lost contacts with the corresponding RM-system, or should possess some additional strategies against RM-systems. In the latter case, the aim of decreasing the number of RM-sites is to promote success of other strategies, rather than to guarantee phage survival independently.

Acknowledgements: the work was supported by RFBR, grants 06–04–49558 and 07–04–91560, and INTAS, grant 05–1000008–8028.

INCORPORATING DIFFERENT TYPES OF EXPERIMENTAL DATA ON DNA-PROTEIN BINDING INTO THE SINGLE *IN SILICO* MODEL

Kulakovskiy I.V.[1, 2]*, ***Favorov A.V.***[2, 3], ***Makeev V.J.***[2, 1]
[1] Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow, Russia,
[2] Institute of Genetics and Selection of Industrial Microorganisms, FGUP GosNIIgenetika, Moscow, Russia,
[3] The Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins, Baltimore, MD, USA
e-mail: ikulakovsky@inbox.ru
* Corresponding author

Motivation and Aim: Genome wide location of transcription factor binding sites (TFBS) at ChIP-chip tiled arrays and even footprints can bring about rather extended DNA segments, which makes a challenge of binding motif identification with traditional techniques. On the other hand the data on protein binding to DNA is available from many different sources of experimental information. Simultaneous analysis of data obtained from such sources as SELEX, ChIP-chip, footprints etc can result in a much clearer signal for DNA-protein binding than usage any of the data sources alone. For instance, the oligos yielded by SELEX strictly correspond to the binding protein, but they are usually short and in practice the binding motif is often distorted. At the same time ChIP-chip arrays give functional binding motifs, often *in vivo*, but the resulting sequences are long and can contain binding signals for proteins different from the test one. Our objective was to make an integrated tool for incorporating different types of experimental data into the single protein binding model.

Methods and Algorithms: For a binding model we have selected the Positional Weight Matrix (PWM) which is traditional motif model for transcription factor binding sites (TFBS) at DNA. The core of the algorithm is SeSiMCMC Gibbs sampler which is used to construct the anchored optimal multiple local alignment (MLA) of raw sequence data. "The anchored" means that any sequence included into MLA should overlap with the anchor sequence initially seeded into the data. This layout allowed incorporating the data of ChIP-chip and SELEX simultaneously. SELEX data was used to place anchors in ChIP-chip sequences. The resulting MLA corresponds to the binding signal for the correct protein.

Results: We paid a particular attention to identify the length of a binding signal, the problem, which is not solved in many signal identification tools. We have tested our system for several TFBS of Human and Drosophila fly and resulting motif models have better selectivity than those built using one source of experimental data.

Conclusion: We created a tool designed to construct a binding motif model from different types of experimental data on DNA-protein binding. Now we can map specific site occurrences at genome sequences within mapped ChIP-chip resulting regions. We can detect genome wide putative TFBS rich regions, which were not covered by ChIP-chip results. This opens a view to compare ChIP-chip results obtained in different experimental environment and study tissue-specific gene expression.

Availability: The source code is available by request. Web-based version of the software tool is planned for release.

COMPUTATIONAL MODEL OF IMPROVING THE EFFICACY OF DUTASTERIDE DRUG

Kushwaha S., Singh P., Shakya M.,* Pardasani K.R.[1]
Deptt. of Bioinformatics, MANIT Bhopal-462051, India
Kushwaha.sandeep04@gmail.com, madhvishakya@yahoo.co.in
[1]Deptt. of Mathematics, MANIT Bhopal-462051, India
kamalrajp@hotmail.com

In the present study efforts have been made to identify new candidate compounds for existing drug to improve efficacy by optimizing various parameters. Single change is considered in compared to multichange, for analog generation. Objective and subjective both approaches were used in identification of descriptors to explore the efficacy of drug dutasteride. A comparison of the calculated binding affinities for structurally similar inhibitors of dutasteride gave suitable analogues. Total 469 analogs are generated and eight are selected for comparative study. To explore the efficacy of drug dutasteride, various softwares and online tools are used in order to annotate the theoretical calculations of analog compounds and drug. In the present study three dutasteride drug analogs showing theoretically more superior results have been predicted. Dutasteride analog with CH_2NH_2, CH_2OH, CF_2OH is identified as the most suitable analogs in the present study that needs to be further evaluated in laboratory.

MODELING AND PREDICTION OF DNA-PROTEIN INTERACTION EVENTS OF TRANSCRIPTION FACTORS (TF) IN CHIP-SEQ EXPERIMENTS

Kuznetsov V.A.*, Singh O., Huck Ng, Wei C.L.
Bioinformatics Institute, 30 Biopolis Street, #07-01 Matrix, Singapore, 138671
e-mail: vladimirk@bii.a-star.edu.sg
* Corresponding author

Motivation and Aim: The next-generation sequencing technologies, capable of producing tens of millions of sequence reads during each instrument run, can be quickly applied to answer to many important genome-wide questions. Recently, Robertson and colleagues (2007) described such an application for identification of binding sites (BSs) of STAT1 TF, using a next-generation sequencing platform (ChIP-seq method). The objective of our work is to develop a computational approach which uses ChIP-Seq data and statistical modeling to identify reliable Transcription Factor Binding Sites (TFBS) for TF in a ChIP-Seq library and predict the total number of the TF BSs in the genome.

Methods and Algorithms: We used two models of statistical distribution of binding events (number of sequence overlap peaks) in ChIP-seq experiments. The first model is the Generalized Discrete Pareto Function and the second model is the Kolmogorov-Waring (K-W) function (Kuznetsov et al, 2007).

Results and Conclusion: Here we present goodness-of fit method to estimatie the number of specific TF binding sites, as well as specificity and sensitivity of ChIPSeq data. The both of our model functions allows us to fit empirical distributions well and thus to extrapolate the functions into high-noisy region (Fig.1). However, we conclude that K-W model better fits to ChIP-seq data (11 libraries) and allows us to estimate fraction of non-observed TFBSs and total number of specific (physical) BSs in genome. We show that our approach in combination with motif search procedures and expression data not only predicts bona fide specific TF BSs, it could also establish practical basis for further optimization of ChIP-Seq protocols.

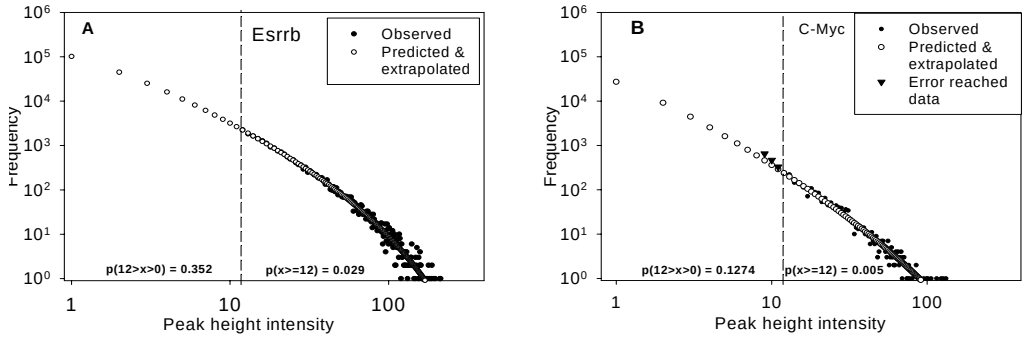


Figure 1. Goodness-of fit analysis of statistics of binding events for TFs (A) ESRRB and (B) c-Myc in mouse genome. Vertical line: cut-off value of reliable events. By K-W model, 62% of low-avidity ESRRB TFBSs and 86% of low-avidity c-Myc TFBS were not represented in the libraries. Thus, vast majority of BSs were not reliably detected, but could be mathematically predicted (on empty circles on the left side of cut-off value).

SURVIVAL SIGNIFICANT AND LOW-DIMENSION GENE SIGNATURES OF BREAST CANCER

Kuznetsov V.A.*, Motakis E.
Bioinformatics Institute, 30 Biopolis Street, #07-01 Matrix, Singapore, 138671
e-mail: vladimirk@bii.a-star.edu.sg
* Corresponding author

Motivation and Aim: Histological grading of breast cancer provides clinically important prognostic information and defines morphological subtypes, informative of patient risk. Approximately 50% of all breast cancers are classified as "Grade 2", which is less informative for clinical decisions than Grade 1 and Grade 3 due to biological heterogeneity of this group's patients and their intermediate risk of cancer recurrence. We have found several small gene signatures (one 5-gene signature and two 7-gene signatures) of histologic breast cancer grades. In this work, we exam the clinical significance of the genes, gene pairs included in small histologic grade signatures.

Methods and Algorithms: A novel statistically-based patients' grouping algorithm called Data-Driven grouping (DDg), is proposed for selection of individual genes and synergetic gene pairs significantly associated with time of patient survival. DDg is based on the semi-parametric Cox proportional hazard regression model which we use in microarray analysis of primary breast cancer samples to identify the most survival significant genes/gene-pairs (by Wald statistics) and to discriminate patients into groups with low and high risk of disease progression. We use 410 microarray U133A&B data of two large cohorts (Stockholm & Uppsala) of breast cancer patients (Stockholm: NCBI dataset label GSE4922; Uppsala: NCBI dataset label GSE1456).

Results: Almost all genes of our three gene signatures were highly significant for poor and good prognosis of disease relapse. Among our findings there are genes that exhibit very high survival significance (e.g. PRC1, MELK, TPX2; $p<1$ E-5) and several transcripts that have not been considered previously as significant breast cancer prognostic markers (FLJ11029 splice variant, C6orf173). We also found several high survival significant synergic gene pairs across genes of our signatures. Finally, we identified functions, networks and pathways of the genes associated with grouping of breast cancer patients based on our 3 molecular signatures.

Conclusion: Our analysis shed light on underlined biological mechanisms of low- and high-aggressive human breast cancer phenotypes and recapitulates our recent re-classification of "Grade 2" breast tumors into "Grade 1-like" and "Grade 3-like" genetic subtypes. Our DDg algorithm can be further used for prediction of new anti-cancer therapeutically perspective genes, their co-regulated partners and disease-related cellular subtypes.

ANCESTRAL ARCHITECTURE OF THE HUMAN CHROMOSOME 17 SYNTENY GROUP IS NON-RANDOMLY MAINTAINED IN MOUSE CHROMOSOME 11

Larkin D.M.[1, 2], Tarasova M.V.[3], Zhdanova N.S.[1]*

[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
[2] Laboratory of Mammalian Genome Biology, University of Illinois, Urbana, IL
[3] Novosibirsk State University, Novosibirsk, Russia
e-mail; zhdanova@bionet.nsc.ru
* Corresponding author N.S. Zhdanova

Motivation and Aim: Comparative mapping of mammalian chromosomes revealed synteny of a number of homologous genes maintained in mammalian evolution. One of the largest autosomic synteny blocks is the region represented by the whole human chr 17 (HSA17). With the exception in a few primate and carnivore species HSA17 genes are linked in the genomes of all placental mammals studied. To elucidate the ancestral structure of this block we performed multi-species comparison of the homologous chromosomes for the representatives of four orders of placental mammals: Primate (human, chimpanzee, macaque), Artiodactyla (cattle), Carnivora (dog, american mink) and Rodentia (mouse).

Methods and Algorithms: Homo sapience (Build 36.1), *Pan troglodytes* (Build 2.1), *Macaca mullata* (Build 1.0), *Canis familiaris* (Build 2.1), and *Bos taurus* (Build 4.0) genes with single known ortholog in the *Mus musculus* (Build 37.0) chr 11 (MMU11) were extracted using Ensemble (v.48; http://www.ensembl.org/). For the *Mustela vison*, the data on the FISH localization for 19 cattle BACs with known orthologous coordinates in the HSA17 and MMU11 were used. We applied the combination of a manual approach together with the SyntenyTracker automated definition of the homologous synteny blocks (HSBs) of genes in the pair-wise comparison for each of the target genomes against the reference mouse genome. The Evolution Highway tool was used to visualize the HSBs and to determine and classify the evolutionary breakpoints (EBs) in each mammalian lineage.

Results: We revealed 9 primate-specific EBs. Ten hominid specific EBs and one macaque specific EBs appeared after the split of primates, one EB is specific to the human lineage, another one to the macaque, and two specific to chimpanzee. Two EB were found in the cattle chromosome 19, five in the dog orthologous chromosomes, and none in the mink when compared to the region of MMU11 orthologous to HSA17 (MMU11: 59,589,725 - 120,852,647 bp). In MMU11 the region orthologous to HSA17 consists of four HSBs with the largest one of 59.1 Mbp (the third largest HSB found in the mouse genome) and a small region of 2.0 Mbp populated with four mouse-specific EBs. The regions homologous to HSA1,5,6,2,7, and 22 constitute the proximal part of MMU11 with the total length of the chromosome 120,9 Mbp and are populated with 12 additional mouse-specific EBs. The distribution of mouse-specific HSB sizes along MMU11 is deviated from the Poisson distribution (p-value < 0.006) suggesting the non-random distribution of the EBs in MMU11 evolution.

Conclusion: Despite the overall high level of the chromosomal rearrangements found in murine rodent lineage (2) the mouse genome maintained the ancestral boreoeutherian architecture of the distal part of MMU11 with the evolutionary breakpoints non-randomly distributed across the mouse chromosome.

MODELING CELL AUTONOMISM ORIGIN IN PROKARYOTES USING EVOLUTIONARY CONSTRUCTOR PROGRAM

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

Motivation and Aim: Progressive evolution involves an organism's complication and its metabolism intensification. Whence and evolutionary ways of complex life forms on the Earth are the significant scientific problems. Many studies using complete genome data of eubacteria, archaebacteria and eukaryotes go in evidence that lots of eukaryotic domains originated rather from bacteria than from mitochondria and plastids. Hence, a eukaryotic cell has been formed owing to autonomism of a member of complex syntrophic prokaryotic community via its self-closure of base regulatory pathways. As an experimental study of evolution is too complex, mathematical modeling becomes the main tool used;

Methods and Algorithms: To model the origin of cell autonomism the previously developed software package "Evolutionary Constructor" (EC) [1] was used. EC is designed for modeling of coevolution in trophically linked populations of haploid organisms. It combines imitation and generalized modeling approaches which allow changing structure of modeled system immediately during the calculation process. In turn it allows us to model such evolutionary processes as mutation and horizontal transfer of gene material (HT);

Results: We have studied long-term evolution of trophic system being represented at starting time as closed "ring" of populations, feeding each other. Each population utilized only one specific (for that population) substrate and produced only one specific (for that population) product. During the calculation a HT might occur with a certain probability. HT could lead to beginnings of novel populations which could utilize more substrates and/or produce more products. It has been shown that in long view the ultimate advantage (the highest population size) obtained the populations having "most complete genome" (i.e. populations utilizing and producing maximal number of substrates/products in given trophic system) or populations having "almost complete genome". In long view such populations eliminated all other members from trophic system while the common metabolism rate grew up.

Conclusion: It was shown that HT in the stable environmental conditions transforms common metabolism of trophic ring to inner metabolisms of separate species becoming autonomous. Selection supplies autonomism, but the novel populations cause trophic ring disintegration to a set of autonomous populations. At that time common metabolism rate grows up. Consequently the criteria of progressive evolution are presented here.

Acknowledgements: Work was supported by the RFBR (No.06-04-49556), RAS Presidium program "Biosphere origin and evolution", project "System Biology: computer-experimental approaches" of RAS Presidium program of molecular and cell biology.

References:

1. S.A. Lashin, V.V. Suslov, N.A. Kolchanov, Yu.G. Matushkin (2007) Simulation of coevolution in community by using the "Evolutionary Constructor" program. *In Silico Biology*, **7:** 261-275.

APPROVAL METHOD FOR TRANSCRIPTION TERMINATION SITES PREDICTION IN *FIRMICUTES*

Lashin S.A.[1]*, Matushkin Yu.G.[1], Khlebodarova T.M.[1], Likhoshvai V.A.[1, 2]
[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
[2] Novosibirsk State University, Novosibirsk, Russia
e-mail: lashin@bionet.nsc.ru
* Corresponding author

Motivation and Aim: Prediction of the sites of transcription termination is the actual bioinformatics problem. It allows primary marking of the genome operon structure. Moreover, in prokaryotes the process of transcription termination is the important stage of gene expression regulation;

Methods and Algorithms: The complete nucleotide sequences of five *Frimicutes* annotated in GenBank (www.gl.iit.edu/frame/genbank.htm) were used for analysis and termination sites prediction. To recognize transcription terminators we have developed original algorithm evolving methods suggested in [1, 2]. As the quality measure of potential terminator we considered the value of discriminant $d = T + G + C + P$, where T is the score value of thymine stretch, G is the Gibbs energy of stem-loop formation (for its calculation the UNAFold package [3] was used), C is the length-penalty of the secondary structure, P is the penalty on the gap between T-stretch and stem-loop structure;

Results: In the study we suggest a novel method of transcription terminators recognition. In comparison with method presented in [2] our algorithm has less beta error (verified using experimentally revealed terminators of *B.subtilis* from [2]). The maps of potential transcription terminators were constructed for five mycoplasms: *Mycoplasma mycoides* (1068 genes), *Mycoplasma mycoides subsp. mycoides* SC Type Strain PG1 (1052 genes), *Mycoplasma hyopneumoniae* J (698 genes), *Mycoplasma hyopneumoniae* 7448 (696 genes), *Mycoplasma genitalium* (515 genes). The rate of fall of transcription terminators number in relation of genome length decreasing is studied.

Conclusion: Specific quantity of recognized transcription terminators was shown to decrease in more high gear than genome size. It can be interpreted as relative growth of the operons length in such organisms in which genome length was decreasing during evolution.

Acknowledgements: Work was supported by the RFBR (No.06-04-49556), RAS Presidium program "Biosphere origin and evolution", project "System Biology: computer-experimental approaches" of RAS Presidium program of molecular and cell biology.

References:

1. d'Aubenton Carafa Y., Brody E., Thermes C. (1990) Prediction of Rh-independent *Escherichia coli* Transcription Terminators. *J. Mol. Biol.,* **216**: 835-858.
2. de Hoon M.J.L., Makita Y., Nakai K., Miyano S. (2005) Prediction of Transcriptional Terminators in *Bacillus subtilis* and Related Species. *PLoS. Comp. Biol.,* **1:** 212-221.
3. R. A. Dimitrov, M. Zuker (2004) Prediction of hybridization and melting for double-stranded nucleic acids. *Biophys. J.*, **87:** 215-226.

MODELING THE TIME-DEPENDENT AND TIME-INDEPENDENT MUTATIONS IN PROKARYOTE CELLS UNDER HEAVY CONDITIONS USING EVOLUTIONARY CONSTRUCTOR PROGRAM

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

Motivation and Aim: Relation between stability and evolvability in biological communities is the key issue of biosphere study. Mathematical models with fixed stoichiometric constants for each substrate are traditionally used to describe ecosystem. Set of those constants specifies conditions of species existence which accords with Liebig's law. Meanwhile in nature the adaptive compensation of utilized substrates (Rubel's law) is observed. In the study the comparative modeling of stability and evolvability of trophically closed communities with metabolism of Liebig and Rubel is investigated;

Methods and Algorithms: For modeling the software package "Evolutionary Constructor" (EC) [1] was used;

Results: Influence of mutations on population survival was modeled under conditions of sublethal deficiency of some substrates. The populations were grouped into trophic ring-like networks (TRLN). Each population utilized specific substrate, which was secreted by its previous TRLN neighbor and produced and secreted specific product that could be utilized by its next TRLN neighbor (as neighbor's specific substrate). Also all populations were needed the same nonspecific substrate injected into system from outside. We have considered two types of TRLN differed in population trophism: 1) insufficiency of some substrates was compensated with redundancy of the other ones (adaptive trophism – TRLN-A); 2) compensation was impossible (Liebig's trophism – TRLN-L). It was shown that if no positive mutations (increasing of substrate utilization efficiency is meant) occurred, it led to the death of both TRLN-A and TRLN-L systems, and they have life time of the same magnitude. Fixation of a positive mutation in population-member of TRLN-L led to a short-time size growth for all populations-members of the network. At the same time limiting regime of functioning did not changed – life time was grown a little. On the other hand, fixation of a positive mutation in member of TRLN-A either significantly (by order of magnitude) increased system's life time or even prevented its extinction;

Conclusion: Modeling has shown TRLN-A to be superior to TRLN-L both in terms of stability and evolvability. Positive mutations, fixed even in one of populations either significantly increased life time of the whole system (additional chance to outlive the "starvation") or even "pulled through" extinction a TRLN-A due to metabolism optimization.

Acknowledgements: Work was supported by the RFBR (No.06-04-49556), RAS Presidium program "Biosphere origin and evolution", project "System Biology: computer-experimental approaches" of RAS Presidium program of molecular and cell biology.

References:

1. S.A. Lashin et al. This issue. Modeling cell autonomism origin in prokaryotes using evolutionary constructor program.

NONCODING RNAS: COUPLERS OF ANALOG AND DIGITAL INFORMATION IN CNS GENE EXPRESSION REGULATION

Laurent, III G.St.[1, 4*], Kanakabandi K.[1], Schtokalo D.[3], Vorobiev D.[3], Faghihi M.[2], Wahlestedt C.[2]

[1] The George Washington University Medical Center, Department of Biochemistry and Molecular Biology, 2300 I Street, NW, Ross Hall 232, Washington, D.C. 20037, USA;
[2] The Scripps Research Institute, Molecular and Integrative Neurosciences Department, 5353 Parkside Drive, Jupiter, FL 33458, USA;
[3] Biorainbow Ltd., Novosibirsk, Russia;
[4] Immunovirology – Biogenesis Group, University of Antioquia, A.A. 1226, Medellín, Colombia
e-mail: gsl@gwu.edu
[*] Corresponding author

The mammalian nervous system expresses numerous noncoding RNAs (ncRNAs). We propose that ncRNAs may be capable of coupling the digital information universe of nucleic acids to the analog universe of cellular protein interactions. First, ncRNAs would allow efficient coupling of energy to information, wherein less energy is required to represent and process more information, condensed in analog and digital form, into smaller spatial and temporal domains, ideal for the environments found in neural tissues. Second, ncRNAs would permit the rapid acquisition of information from the environment, along with the rapid flexible processing and elimination of that information when no longer necessary. Third, ncRNAs would facilitate accelerated evolution of an organism's information content and functional computational systems. These features allow ncRNAs to form complex circuitry for the precise control of gene expression in high complexity systems such as mammalian stress response and synaptic plasticity. For example, we have identified stress response functions for the highly conserved noncoding antisense transcript for β-secretase-1 (BACE1), a critical enzyme in Alzheimer's disease (AD) pathophysiology. This ncRNA (BACE1-AS) concordantly regulates BACE1 mRNA and subsequently BACE1 protein expression in vitro and in vivo, using both analog and digital information coded within its 2kb sequence.

COMPUTER SYSTEM FOR MODELING AND ANALYSIS OF PLANT TISSUE GROWTH AND DEVELOPMENT

Lavreha V.V.[1]*, Penenko A.V.[2], Nikolaev S.V.[1], 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: vvl@bionet.nsc.ru
* Corresponding author

Motivation and Aim: Software development for analysis and modeling of plant tissue a very important problem in developmental biology. Growth and development of tissue are determined by cellular biomechanics, molecule transportation, and cellular responses to their microenvironments. Biologically motivated models of such processes, when incorporated, would help to study the influence of these processes on plant tissue development. Explicit account of geometric characteristics of biological object allows carrying out quantitative analysis of mechanisms of growth and development regulation.

Comparing models with real observed tissue structures is the very important step in model verification with experimental data and its interpretation. The image processing is one of the powerful instruments for such an analysis. As the result of image processing the geometric model of real object (cell ensemble and/or tissue) could be obtained. That is why such an approach to the building of the model might be used as "morphology module" of the model development system.

Results: Presented system was designed to support the modeling process of cell ensemble development, i.e. constructing geometry of biological objects based on real images, editing its biological and biophysical characteristics, and simulating of its functional state dynamics.

This system can be used for studying influence of biophysics and molecular-biological processes on morphogenesis both separately and in the aggregate; and for simulating growth and morphogenesis of 'planar' plant tissues.

Our computer system has module architecture that allows extending easily its functionality. Now it includes following modules: (1) module of cells growth and division simulation with the following possibility. The regulation of morphogens expression and its influence on cell functioning and differentiation have been taken into account; (2) the module calculating water, morphogens and "nutrients" transport through the tissue; (3) the module calculating tissue deformation based on biomechanics models; (4) the module acquiring cell ensemble structure from images.

Conclusion: Main aspects of modeled objects are: orientation to plant tissue (without migration of cells); mechanic properties of plant cells are defined by cell walls; the diffusion flux depend on square of the cell wall; arbitrary form of cell ensemble is possible. Functional aspects include: the graphic user interface; the library of models; reconstruction of the cell ensemble geometry from the biological objects images.

Acknowledgements: This work was supported by the RFBR grant: 08-04-01214-a "Mathematical modeling and analysis of structure homeostasis mechanism of the stem cell niche in Arabidopsis thaliana shoot apical meristem".

GENETIC DIVERSITY OF TRITICUM AESTIVUM L. ON POWDERY MILDEW RESISTANCE (*BLUMERIA GRAMINIS* DC. F. SP. *TRITICI* GOLOVIN)

Lebedeva T., Zuev E.*
N.I. Vavilov All-Russian Research Institute of Plant Industry, St. Petersburg, Russia
e-mail: e.zuev@vir.nw.ru
* Correspondimg author

Motivation and Aim: Powdery mildew caused by *B. graminis* f. sp. *tritici* is a widespread foliar disease of common wheat in regions with cool, wet conditions. At the present day genes for powdery mildew resistance (*Pm* genes) have been described at 34 gene loci in common wheat. In the wheat breeding much less genes are used. High resistance of new wheat varieties is limited by their time-life because of continuous appearance of fungus biotypes with new virulence. That is why the research of new resistant forms is of a great importance.

Methods and Algorithms: Our study is devoted to estimation of resistance potential of common wheat. The estimation of resistance of wheat varieties from VIR collection was made in the field (adult plants) and laboratory (seedlings). The field population of fungus had virulent to *Pm1, Pm2, Pm4, Pm6, Pm8, Pm9, Pm10, Pm10+15, Pm11, Pm19*. The test lines with *Pm3(a,b,c,d), pm5, pm7* and *Pm18* had strong leaf chlorosis. Thee test-line Wembley 14.31 (*Pm12* gene from *Aegilops speltoides*) and test-line BR 93N (*Pm16* gene from *Triticum dicoccoides*) had immunity to fungus field population. At the seedling stage only test-lines with *Pm12* and *Pm16* had immunity, the others were susceptible to the disease.

Results and Conclusion: The next varieties shown high resistance at the adult stage: k-64544, k-64555 (Russia), k-610694, 610595, k-610700 (Ukraine), k-64480 (Poland), k-611828 (Czech R.), k-610705 (Serbia), k-141951 (Germany), k-141399, k-61059, k-64433, k-64434, k-64435, 64436 (Sweden). Varieties k-604655, 604674 (Canada), k-141392, k-64554 (Russia), k-610709, k-610710 (Czech R.) were intermediate. Varieties from Estonia Helle (k-610818) and Meri (k-610819) had strong leaf chlorosis. At the seedling stage high resistance was shown by only Sweden varieties and Brasil line BR34 (k-62165). The resistance of the Sweden varieties is likely to be controlled by Pm12 or Pm16 or a new gene. The high resistance to powdery mildew of Brasil line BR34 was established to be controlled by one dominant gene non allelic *Pm12* or *Pm16*.

NUCLEOSOME FORMATION POTENTIAL ESTIMATION VIA DINUCLEOTIDE PERIODICITY PREFERENCES

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

Motivation and Aim: Currently theoretical prediction of DNA sequences possessing higher nucleosome formation efficiency are still far from wide genome application.

Methods and Algorithms: The PHASE method measures the abundance of phased dinucleotides in DNA sequence implying the double helix periodicities (~10.5 bp per turn). Any observed dinucleotide considered either as a phased or not depending on the number of other dinucleotides of the same type anywhere one turn apart in both directions of DNA strand. The method compared the probabilities of specifically phased dinucleotides with those expected by chance (1^{st} order markov model). Total score is constructed on the basis of dinucleotide weight matrix according to NLG approach [1], but instead of every dinucleotide type we have one not phased subtype and two phased ones. Mammal's nucleosome formation sites [2] were used to estimate any specific dinucleotide phasing in a certain site position. The growth of PHASE score implied the better nucleosome formation.

Results: At first we verified that PHASE score ensured correct classification of stable nucleosomes [3] and anti-nucleosomal DNA [4] and splice junctions. The higher and lower scores were found for housekeeping and tissue-specific gene promoters, respectively [5,6]. Similar contrast was also observed for promoters from TRRD [7] expressed in restricted and wide range of tissues. The scan of full-length human chromosomes approved that sites of dinucleotide periodicity overrepresentation are uniformly distributed as against to clustered underrepresentation sites. We demonstrated that DNase I hypersensitivity sites mapped in genome-scale microarrays [8] characterized by decreased PHASE score.

Conclusion: The PHASE model may be successively applied for prediction of nucleosome formation potential of genome DNA.

References:

1. V.G. Levitsky et al., (2007) Effective transcription factor binding site prediction using a combination of optimization, a genetic algorithm and discriminant analysis to capture distant interactions. *BMC Bioinformatics,* **8:** 481.
2. V.G. Levitsky, et al., (2005) NPRD: Nucleosome Positioning Region Database. *Nucl Acids Res*, **33:** 67-70.
3. H.R. Widlund et al., (1997) Identification and characterization of genomic nucleosome-positioning sequences. J Mol Biol, **267**: 807-817.
4. H. Cao et al., (1998) TGGA repeats impair nucleosome formation. *J Mol Biol*, **281**: 253-260.
5. V.G. Levitsky et al., (2001) Nucleosome formation potential of eukaryotic DNA: tools for calculation and promoters analysis. *Bioinformatics*, **17**: 998-1010.
6. M. Ganapathi et al., (2005) Comparative analysis of chromatin landscape in regulatory regions of human housekeeping and tissue specific genes. *BMC Bioinformatics*, **6**: 126.
7. N.A. Kolchanov et al., (2002) Transcription Regulatory Regions Database (TRRD): its status in 2002. *Nucl Acids Res*, **30:** 312-317.
8. P.J. Sabo et al., (2006) Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays. *Nature Methods,* **3**(7): 511-518.

MODELLING AND COMPARATIVE ANALYSIS OF AUXIN TRANSPORT MECHANISMS IN SHOOT AND ROOT

Likhoshvai V.A.[1, 4]*, Akberdin I.R.[1], Mironova V.V.[1], Omelyanchuk N.A.[1], Fadeev S.I.[2, 4], Mjolsness E.[4]

[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
[2] Institute of Mathematics, SB RAS, Novosibirsk, Russia
[3] Institute of Genomics and Bioinformatics, University of California, Irvine, USA
[4] Novosibirsk State University, Novosibirsk, Russia
e-mail: likho@bionet.nsc.ru
* Corresponding author

Motivation and Aim: Plants compared to the most animals do not complete establishment of their body plan in embryogenesis but continually produce new lateral organs in the course of postembryonic development. A general mechanism of new lateral organ initiation is common for both shoot and root apical meristems (SAM and RAM, respectively) and consists of local accumulation of auxin at the margin of stem cell niche [1]. In the case of RAM these local maxima appear on the way of the auxin stream coming from the shoot along the root longitudinal axis. In the SAM local auxin maxima occur in the peripheral zone lying as a ring surrounding the SAM summit. Polar auxin transport machinery producing these local maxima consists of the influx and efflux carriers. Previously we have developed the 1D model (1), for studying of auxin distribution along the longitudinal axis in the plant root [2].

Methods and Algorithms: We converted the model (1) into the model (2r) attempting to present auxin stream in the SAM: we replaced column cell ensemble by the ring cell ensemble, made the auxin stream within the ring bidirectional and excluded from the model (2r) auxin input and dissipation. All other parameters and conditions in the model (2r) were kept the same as in the model (1).

Results: The model (2r) did not give rise to any local maximum of auxin in the ring. We changed the equation according to [3], where regulation of auxin transport from the cell depends on auxin concentration in the neighboring cell. This model (2s) demonstrated local auxin maxima similar to [3].

Conclusion: Numerical simulations of models (1) and (2) showed that each of them is effective in only one case: model (1) in RAM and model (2s) in SAM This provided insights into the difference in auxin maximum formation in SAM and RAM borders related to the difference in the regulation of polar auxin transport. It's allowed to suggest that in the shoot apex there is an auxin absorption mechanism and the cell with highest auxin level works like "vacuum cleaner", but in the root auxin extrusion mechanism is prevalent and each cell works like "spray".

References:

1. De Smet I, Jürgens G. Patterning the axis in plants--auxin in control. (2007) *Curr Opin Genet Dev.*, **4**:337-43.
2. В.А. Лихошвай, Н.А. Омельянчук, В.В. Миронова, С.И. Фадеев, Э.Д. Мелснесс, Н.А. Колчанов Математическая модель паттерна распределения ауксина в корне растения (2007) *Онтогенез*, **6**: 446-456.
3. Jonsson et al. (2006) An auxin-driven polarized transport model for phyllotaxis *PNAS*, **103**: 1633–1638.

MODELLING OF AUXIN CONTROL OF ROOT PATTERNING

Likhoshvai V.A.[1, 4]*, Omelyanchuk N.A.[1], Mironova V.V.[1], Fadeev S.I.[2, 4], Yosiphon G.[3], Mjolsness E.[3]

[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
[2] Institute of Mathematics, SB RAS, Novosibirsk, Russia
[3] Institute of Genomics and Bioinformatics, University of California, Irvine, USA
[4] Novosibirsk State University, Novosibirsk, Russia
e-mail: likho@bionet.nsc.ru
* Corresponding author

Motivation and Aim: There are two major types of root systems in the flowering plants: fibrous and taproot. Most monocots have fibrous root, where primary root ceases to elongate and root system forms from adventitious roots arising from hypocotyl. A taproot mainly attributed to dicots is formed when the primary root becomes the central and most important feature of the root system. In both lateral root formation takes place. The following three types of auxin concentration maxima were experimentally detected in plants. The terminal auxin maximum (TAM) is always located in the root tip [1]. The internal auxin maxima (IAM) are temporally observed in the provascular tissue preceding the initiation of lateral root meristems [2]. As well as a hypocotyl auxin maximuma (HAM) precede to adventitious root meristems. In the course of plant development the auxin flow from the shoot is increased following to the shoots growth.

Methods and Algorithms: We analyzed the influence of increasing values of auxin flow from the shoot on solutions of 1Da [1] and 1Db [3] models of auxin distribution in root. The difference is that 1Db model accounts for cell divisions in the root meristem. 1Da deals with the fixed cell number. The "tolerant" set of parameters fit the curve of auxin distribution in the arabidopsis root. The "sensitive" set was adjusted from the "tolerant" one by increasing the efficiency of auxin transport inhibition by auxin.

Results: 1Db model with slowly increase of the auxin flow reproduces maintaining of the TAM position and reiterative appearance of IAM at the both sets of parameters. The position of TAM was permanent for the "tolerant" and variable in some interval for the "sensitive" set of parameters. With the "sensitive" set of parameters, 1Db demonstrated more than one IAM as well as HAM formation. In the 1Da model with the «tolerant» set of parameters the TAM shifted proximally, further increase caused oscillations in auxin concentration in the root end. Proximal shift of the TAM was also the first response in 1Da with "sensitive" parameters, following by appearance of IAM and shifting back the TAM. Finally, the TAM disappeared and oscillations of auxin concentration were recorded in the basal part of the axis corresponding to hypocotyl.

Conclusion: In our model increasing auxin flow resulted in appearance of proximal auxin maxima. These maxima in vivo precede the initiation of lateral and adventitious roots. This fact and different model behavior at two sets of parameters indicated that the difference in the auxin acropetal transport regulation may be the main factor responsible for plant diversity in root architecture.

References:

1. V.A. Likhoshvai et al. (2007) Mathematical model of auxin distribution in the plant root, *Ontogenez,* V. 38 N. 6 P. 374–382.
2. de Smet et al. (2007) Auxin-dependent regulation of lateral root positioning in the basal meristem of Arabidopsis*, Development*, 134, 681-690.
3. V.V. Mironova et al. (2008) Auxin regulation of its own transport determines the root tip structure in plants, *In this press*.

GENETIC CONSTRUCTOR: A COMPUTER RESOURCE FOR MOLECULAR-GENETIC SYSTEMS MODELING

Likhoshvai V.A.*[1, 2], Tikunova N.V.[1, 2], Kachko A.V.[1, 2], Khlebodarova T.M.[1]
[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
[2] Novosibirsk State University, Novosibirsk, Russia
e-mail: likho@bionet.nsc.ru
* Corresponding author

Motivation and Aim: Study of patterns of nature and artificial gene nets is an important problem of modern biology today. This problem could be solved by computer recourses that allow modeling of functions of genetic systems, analysis of their properties using *in silico* experiment, and prediction of new knowledge about the genetic systems. This study is devoted to development of the resource based on an original unified standard of model specification (SiBML standard).

Methods: "Genetic constructor" was designed on the base of computer software MGSmodeller and modeling resources of SiBML standart [1, 2], which allowed development of elementary mathematical models irrespective of other models, but provide their compatibility.

Results: A computer resource "Genetic constructor" is an integrate system including data base of the elementary subsystem models [3], elementary models of phage λ ontogenesis and engineering environment using genetic maps. "Genetic constructor" allows modeling of the regulatory circuits of genetic system and studies their functioning in various conditions of external and internal environments. The potentialities inherent to "Genetic constructor" were demonstrated on the example of Elowitz-Leibler repressilator based on a synthetic oscillatory network of the *lacI, λ cI* and λ *tetR* genes [4] and on the example of genosensor based on the *E. coli yfiA* gene [5].

Conclusion: The application of "Genetic constructor" for analysis of Elowitz-Leibler repressilator allowed not only to reproduce behavior of the repressilator in the experiment [3] but to demonstrate new interpretations and characteristics of the process. Use of "Genetic constructor" for the analysis of genosensor functioning let us to estimate the peculiarity of the regulatory mechanisms of unknown promoter *of* the *E. coli yfiA* gene.

References:

1. V.A. Likhoshvai, A.V. Ratushny (2007) Generalized hill function method for modeling molecular processes. *J. Bioinform. Comput. Biol.*, 5: 593-610.
2. N.L. Podkolodny et al. (2006) An integration of the descriptions of gene networks and their models in sigmoid (cellerator) and genenet. In: *Proceedings of 5th BGRS,* 3: 86-90.
3. A.V. Ratushny et al. (2006) Database of mathematical models of molecular genetic processes (ModelER). *Certificate of authorship RF* №2006620196.
4. M. Elowitz, S. Leibler (2000) A synthetic oscillatiry network of transcriptional regulators. *Nature,* **403:** 335-338.
5. N.V. Tikunova et al. (2007) A computational-experimental approach to designing a polyfunctional genosensor derived from the *Escherichia coli* gene *yfiA* promoter. *Dokl. Biochem. Biophys.*, **417**: 357-361.

RNA STRUCTURES UPSTREAM THE 2-ISOPROPYLMALATE SYNTHASE ENCODING GENE IN α-PROTEOBACTERIA AND ACTINOBACTERIA

Lopatovskaya K.V., Seliverstov A.V., Lyubetsky V.A.*
Institute for Information Transmission Problems RAS (Kharkevich institute), Moscow, Russia
e-mail: lyubetsk@iitp.ru
* Corresponding author

Motivation and Aim: In many proteobacteria the leucine synthesis is regulated by classical attenuation, while in most α-proteobacteria it is apparently not the case for gene *leuA* encoding 2-isopropylmalate synthase. Regulation of gene *leuA* is studied.

Methods and Algorithms: Previously published algorithms are employed to analyze all complete genomes of α-proteobacteria from GenBank (NCBI).

Results: In many α-proteobacteria (Rhizobiales: *Agrobacterium tumefaciens*, *Aurantimonas* sp. SI85-9A1, *Brucella* spp., *Fulvimarina pelagi*, *Mesorhizobium* spp., *Rhizobium* spp., *Sinorhizobium* spp.; Rhodospirillales: *Magnetospirillum* spp.; Rhodobacterales: *Dinoroseobacter shibae*, *Jannaschia* sp. CCS1, *Loktanella vestfoldensis*, *Oceanicola* spp., *Rhodobacterales bacterium* HTCC2654, *Rhodobacter* spp., *Roseobacter denitrificans*, *Roseovarius* spp., *Sulfitobacter* spp., *Alpha proteobacterium* HTCC2255) a conserved pseudoknot with long shoulders and a leader peptide gene with a run of leu-codons ("LEU-regulation") were found to precede a *single paralog* of gene *leuA*. In Rhizobiales LEU-regulation structure contains an additional variable helix within the conserved pseudoknot. In Rhodospirillales and in most Rhodobacterales the pseudoknot is located at a short distance from the start of gene *leuA*; in other cases it is located further away and separated by a long low-conserved region, thus precluding interaction between the pseudoknot and the ribosome binding site. In all referred cases the attenuation-characteristic terminator is lacking from the entire region between the leader peptide gene and gene *leuA*. Neither of *leuA* paralogs having LEU-regulation is located close to other *leu* or *ilv* genes. Thus, in *Magnetospirillum* spp. such paralog is essentially more distant in terms of sequence similarity from gene *leuA* in *E.coli* than from another *leuA* paralog in *Magnetospirillum* spp., which is positionally coupled with *ilv* genes and undergoes classic attenuation regulation with *ilv* genes. Therefore, one might suggest that *leuA* gene with its LEU-regulation was horizontally transferred in α-proteobacteria, while classic attenuation existed in the common ancestor of α- and γ-proteobacteria.

Conclusion: LEU-regulation is related to the ortholog complex of isopropylmalate/homocitrate/citramalate synthases, which lacks the *leuA* representative in *E.coli* and in most proteobacteria. Also, Rhizobiaceae possess a single *leuA* gene with LEU-regulation, which indicates its functional activity. Thus, the *leuA* gene in Rhizobiaceae can be hypothesized to be a xenolog. Another conserved RNA structure containing a pseudoknot and leader peptide gene different from both attenuation and LEU-regulation structures was found in actinobacterial genera *Acidothermus*, *Actinomyces*, *Arthrobacter*, *Brevibacterium*, *Clavibacter*, *Corynebacterium*, *Frankia*, *Kineococcus*, *Leifsonia*, *Mycobacterium*, *Nocardia*, *Rhodococcus*, *Saccharopolyspora*, *Salinispora*, *Streptomyces*, and *Thermobifida*.

MOLECULAR MODEL OF TERTIARY STRUCTURE FOR HUMAN FULL-LENGTH CYTOCHROME P45017α

Lukashevich O.P.*, Gilep A.A., Usanov S.A.
Institute of Bioorganic Chemistry of the National Academy of Sciences of Belarus, Minsk, Belarus.
e-mail: luko@iboch.bas-net.by

Motivation and Aim: Cytochrome P45017α (CYP17α) plays an important role in biosynthesis of steroid hormones [1]. Its three-dimensional structure has not been resolved, because the enzyme is a hydrophobic membrane bound hemeprotein. The only approach to obtain information about 3D structure of CYP17α is a homology modeling.

Methods and Algorithms: The sequence alignment has been performed with ClustalW 2.0 [2]. The initial 3D model of CYP17α has been generated using Modeller 8v2 [3]. The model has been validated with Procheck procedure and chosen for further refinement. Molecular dynamics (MD) simulation has been carried out at 1 atm and 300K with the NPT ensemble.

Results: A blastp search (www.ncbi.nih.gov/blast) confirmed that CYP17α has the sequence similarity 36,7% with CYP1A2 and CYP2R1. The coordinates of N- terminal residues (1-26) in crystal structures of these hemeproteins are missing. Because of its importance for substrate binding we reconstructed the secondary structure of N-terminus. Figure1. displays a schematic representation of the refined homology model for human full-length CYP17α. The volume of binding pocket is 189Å^3. The total energy of the model decreased substantially in the first 250 ps of dynamics simulation and stabilized after 600 ps equilibration. The RMSD of the heavy atoms compared the starting coordinates increased slowly in the first 700 ps time period and then reached a plataue in the sequent simulation time. All the properties converged after 700 ps MD simulation, indicating that the model is stable and can be used for further docking of steroids to elucidate the mechanism of their transformation (hydroxylation).

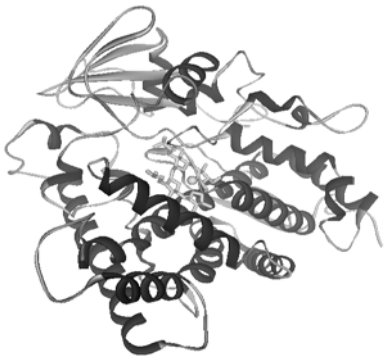


Figure 1. Full-length model of CYP17α

Conclusions: The model for human full-length CYP17α was constructed using homology modeling and MD simulation and critically assessed with stereochemical and energetic aspects. The largest flexibility was found in the preceding loop region of helix A and the region between helices F and G. The latter is considered to be responsible for substrate binding.

References:

1. A.A.Gilep et al. (2003) Molecular cloning and heterologous expression in E. coli of cytochrome P45017alpha. Comparison of structural and functional properties of substrate-specific cytochromes P450 from different species, *Biochemistry (Mosc)*, **68(1):** 86-98.
2. J.D.Thompson et al. (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools, *Nucleic Acids Res*; **25:** 4876-4882.
3. A.Sali, T.L.Blundell (1993) Comparative protein modelling by satisfaction of spatial restraints, *J Mol Biol,*; **234**: 779-815.

OsPAD: A SYSTEMIC PROTEOME ANNOTATION DATABASE FOR *ORYZA SATIVA* 2D-PAGE

Luo C., Chen M.*
Department of Bioinformatics, College of Life Sciences,
Zhejiang University, Hangzhou 310058, China
* Corresponding author: E-mail: mchen@zju.edu.cn

Motivation and Aim: Two-dimensional polyacrylamide gel electrophoresis (2D-PAGE) is a widely used separation technique that can identify target proteins existing in a tissue. It has the power to monitor global changes of protein expressions that occur in the tissues, organisms, and/or under different circumstances. However, 2D-PAGE suffers from the restriction of the following identification methods such as PMF and MS/MS (For instance a low abundant spot is hardly possibly identified.). For rice, until now few systemic 2D-PAGE databases are available, one of which is RPD (Rice Proteome Database). It is an online resource for *Oryza sativa* which contains information on proteins identified from 23 tissues and organelles 2D-PAGE reference maps, nevertheless only about 24.4% of total spots are identified with little correlative information. We are motivated to annotate the remaining spots automatically by using a data mining technique and develop the database OsPAD (*Oryza sativa* Proteome Annotation Database of 2D-PAGE) providing with comprehensive protein information.

Methods and Algorithms: Based on the original RPD data (23 gel images data sets), we integrated relative data from KEGG and Uniprot databases. For each tissue/organelle, candidate proteins were selected from the KEGG pathways which were generated by the identified spots. Then in certain error ranges, we queried these candidate proteins against Uniprot according to the theoretical *pi* and *mw*, the three most accurate hits were choosed. These hits are regarded as the potential proteins for each spot. Using this approach, 9713 out of 12840 gel spots were annotated with Uniprot/KEGG information.

Results: OsPAD users can query *pi*, *mw* and tissue/organelle (selective) to find out potential proteins for a experimental gel spot. They can also search items with Uniprot/KEGG information or certain keywords to get detailed information on the interested protein(s). With community effort, experimental data submission function can help to create a fully open-source 2D-PAGE protein expression data analysis system. The results can be freely downloaded and used for data mining protein expression profiles across sets of 2D-PAGE data from research experiments.

Conclusion: In conclusion, OsPAD is a systemic proteome annotation database for *Oryza sativa* 2D-PAGE with integrated rice proteomics data. The information obtained from the rice proteome using expression-data-mining and comparing will be helpful in predicting the function of the unknown proteins. Thus our database can help biologists in better understanding of the rice development, growth or other physiological activities.

ANALYSIS OF SIG3, SIG4, AND SIG6 EVOLUTION ON THE BASIS OF NEW GENOMIC DATA

Lysenko E.A.[1, 2]*, Seliverstov A.V.[1], Lyubetsky V.A.[1]
[1] Institute for Information Transmission Problems RAS (Kharkevich Institute), Moscow, Russia
[2] Institute of Plant Physiology RAS, Moscow, Russia.
e-mail: genlysenko@mail.ru
* Corresponding author

Motivation and Aim: Sigma subunits are basal transcription factors of bacterial RNA polymerases. However, eukaryotes have such RNA polymerases too: in plastids of different algae and land plants transcription is directed by RNA polymerases of cyanobacterial origin. Plant sigma factors are encoded by a small family of nuclear *Sig* genes. Their evolution was studied previously [1] but some steps are still unclear. We have searched through newly completed plant genomes to clarify the evolution of this gene family in higher plants.

Methods and Algorithms: Standard BLAST procedures and multiple alignment with program AlignW were used for the analysis. Prediction of plastome promoters was based on a clique finding method described earlier [2].

Results: We have searched through completed genomes of the moss *Physcomitrella patens*, dicotyledonous flowering plant *Populus trichocarpa*, monocotyledonous flowering plant *Oryza sativa*, and not completed genomes of some other plants. We have revealed previously unknown genes coding for plant sigma factors: *OsSig3, PpSig2b, PpSig2c, PpSig2d, PtSig1a, PtSig1b, PtSig2a, PtSig2b, PtSig3, PtSig4, PtSig5a, PtSig5b, PtSig6, VvSig4*. Also genomic copy of *ZmSig3* has been found (only cDNA of this gene was known previously). The revealed genes were classified on the basis of multiple alignments, evolutionary tree construction procedure, and comparison of intron positions.

In *Arabidopsis thaliana*, plastome gene *ndhF* is Sig4-dependent [3]. We have analysed 5'-upstream *ndhF* regions from all known plastomes of photosynthesizing flowering plants. The analysis revealed a homologous promoter in all Brassicaceae species (including *A. thaliana*), in representatives of closely related families (*Citrus sinensis, Gossypium sp.*), and in two very distantly related species – *Vitis vinifera* and *Platanus occidentalis*. In the majority of species this region was non-homologous to *A. thaliana* Sig4-dependent promoter.

Conclusion: The data obtained let us clarify timing of evolution of less ancient *Sig* genes (3, 4, and 6) and introduce a new scheme of their evolution: genes *Sig3* and *Sig6* arose after divergence of mosses and vascular plants but before divergence of mono- and dicotyledonous flowering plants. Gene *Sig4*, probably, arose later – after separation of eudicotyledonous plants, but before their divergence into stem and core eudicotyledons.

References:

1. E.A.Lysenko (2006) Analysis of the evolution of the family of the *Sig* genes encoding plant sigma factors, *Russ. J. Plant Phys.,* **53**: 684-694.
2. A.V.Seliverstov, V.A.Lyubetsky (2006) Translation regulation of intron containing genes in chloroplasts, J. Bioinform. Comput. Biol., **4**: 783-790.
3. J.J.Favory et al. (2005) Specific function of a plastid sigma factor for *ndhF* gene transcription, *Nucleic Acids Res.,* **33**: 5991-5999.

A MODEL OF REGULATORY SIGNAL EVOLUTION

Lyubetsky V.*, Zhizhina E., Rubanov L.
Institute for Information Transmission Problems RAS (Kharkevich institute), Moscow, Russia
e-mail: lyubetsk@iitp.ru
* Corresponding author

Motivation and Aim: Evolution of genes, proteins and species was studied for a long time; various models of the evolution have been developed. We aimed to develop an evolution model of regulatory site, specifically the site which performs by folding alternative RNA secondary structures. The site of classic attenuation regulation is an example.

Methods and Algorithms: Given a phylogenetic tree e.g. of species, each leaf is juxtaposed with contemporary site of classic attenuation regulation. The objective is to reconstruct such sites in ancestral nodes of the tree. The configuration is a function which juxtaposes each internal node with a nucleotide sequence. We sought for the configuration that yields the absolute minimum of two items summed. The first item describes primary structure dynamics as per some model of nucleotide change allowing for deletion and insertion. The second item represents a condition of the most possible conservatism of antiterminator-terminator pairs along entire paths from each leaf to the root. These pairs are selected in the sequences corresponding to the tree nodes in accordance with the configuration found. We proposed and implement the algorithm to search the absolute minimum of the items sum.

Results: The algorithm was extensively tried out at artificial and natural data. The following fragment of the result was obtained for regulatory site of threonine biosynthesis in gamma-proteobacteria. Only paths from three leaves are shown; complete result for this example (and others) is given at http://lab6.iitp.ru/docs/anneal/primer.htm along with appropriate trees, species names, etc. The antiterminator is marked by underline, the terminator has grey background. Node name and the second item value is given to the right for each edge and the whole paths from the leaves (species) EO, HI, SON to the root N01 over respective nodes of the tree.

agtcggggcgggctgttgcgtccacgtaactaaacaacgagcccgcatccgaccaggatgcgggcgtttctctc **N01** H3=-40.2
agttggggcgggctgatgcgcccaaaaaattcaacaacgagcccgcatccaacaaagatgcgggctttttctt **N02** H3=-35.5
tgttggggcgggctgatgcgcgcaaaaaattcaaaaaaaagcccgtatccaacaagatgcgggctttttttta **N03** H3=-34.8
taatggtgcgggctgatgcgcacaaaaaattcaaaaaaaagcccgtatccaacaaaatgcgggctttttttta **N04** H3=-41.2
taacggtgcgggctgacgcgtacagaaaacacagaaaaaagcccgcacctgaacagtgcgggctttttttta **N05** H3=-52.0
taacggtgcgggctgacgcgtacagaaaacacagaaaaaagcccgcacctgaacagtgcgggctttttttt **N06** H3=-52.0
taacggtgcgggctgacgcgtacagaaaacacagaaaaaagcccgcacctgaacagtgcgggctttttttt **N08** H3=-42.0
taacggtgcgggctgacgcatacaaagattccagaaaaagcccgcaccgaacagtgcgggctttttttt **EO Σ=-297.9**

agtcggggcgggctgttgcgtccacgtaactaaacaacgagcccgcatccgaccaggatgcgggcgtttctctc **N01** H3=-40.2
agttggggcgggctgatgcgcccaaaaaattcaacaacgagcccgcatccaacaaagatgcgggctttttctt **N02** H3=-35.5
tgttggggcgggctgatgcgcgcaaaaaattcaaaaaaaagcccgtatccaacaagatgcgggctttttttta **N03** H3=-34.8
taatggtgcgggctgatgcgcacaaaaaattcaaaaaaaagcccgtatccaacaaaatgcgggctttttttta **N04** H3=-36.6
aaatggtgcgggttagtgcgaacaaaaagatcaaacgaaaaaacccgtatccaacaaaatgcgggttttttata **N09** H3=-43.6
aatggtgcgggttagtgttagcaaaaacaacaaacgaaaaaaccggtatccaactaaatgcgggttttttata **N10** H3=-14.8
aatggtgcgggttagtgcagcaaaaacaagatacagaaaacccgcgattcaactgaatagcgggttttttata **HI Σ=-205.7**

agtcggggcgggctgttgcgtccacgtaactaaacaacgagcccgcatccgaccaggatgcgggcgtttctctc **N01** H3=-40.2
agttggggcgggctgatgcgcccaaaaaattcaacaacgagcccgcatccaacaaagatgcgggctttttctt **N02** H3=-42.3
agtgggggcgggctgatacaccctaaagaatttaacgacgagcccgcttcccacaaagaagcgggctttttgtt **SON Σ=-82.5**

Conclusion: A model and algorithm are proposed to reconstruct ancestral states of the regulatory signal in classic attenuation regulation.

EVOLUTION OF ANTISENSE TRANSCRIPTS IN VERTEBRATE GENOMES

***Makalowska I.*[1, 2*]*, Lin C-F.*[3]

[1] Institute of Molecular Biology and Biotechnology, Adam Mickiewicz University, Poznan, Poland;

[2] Center for Computational Genomics, Hucks Institutes of Life Sciences, Pennsylvania State University, University Park, USA;

[3] Institute of Bioinformatics, University of Muenster, Muenster, Germany;

e-mail: izabelam@psu.edu

* Corresponding author

Between five and fourteen per cent of genes in the vertebrate genomes do overlap sharing some intronic and/or exonic sequence. It was observed that majority of these overlaps are not conserved among vertebrate lineages. Although several mechanisms have been proposed to explain gene overlap origination the evolutionary basis of these phenomenon are still not well understood. Here, we present results of the comparative analysis of several vertebrate genomes. The purpose of this study was to examine overlapping genes in the context of their evolution and mechanisms leading to their origin.

Based on the presence and arrangement of human overlapping genes orthologs in rodent and fish genomes we developed 15 theoretical scenarios of overlapping genes evolution. Analysis of these theoretical scenarios and close examination of genomic sequences revealed new mechanisms leading to the overlaps evolution and confirmed that many of the vertebrate gene overlaps are not conserved. This study also demonstrates that repetitive elements contribute to the overlapping genes origination and, for the first time, that evolutionary events could lead to the loss of an ancient overlap.

Our study shows that there's no single mechanism responsible for the overlap origination. In principle, any mechanism of a new exon or a new gene origination may lead to a gene overlap. In the light of presented results, we can conclude that the major forces in the overlapping genes evolution are transposition and exaptation - a process that gives rise to new genes or new variants from preexisting nucleotide sequences. Additionally, results of our study imply that origin of overlapping genes is not an issue of saving space and contracting genomes size. Although there are some implications on functional importance of overlapping genes, the present analysis shows that most gene overlaps evolve stochastically, the same way as other genomic features, and without any positive pressure on the overlap presence.

Birth as well as most probably death of gene overlaps occurred over the entire time of vertebrate evolution and there wasn't any rapid origin or 'big bang' in the course of overlapping genes evolution. Majority of gene overlaps are lineage specific and are not conserved among vertebrates and this study demonstrates that in order to fully understand the evolution of overlapping genes one has to study many genomes in minute details. Studies on a limited number of species may lead to false conclusions.

GENOMIC SCRAPYARD OR HOW GENOMES UTILIZE ALL THAT JUNK

***Makalowski W.*[1*]*, Gotea V.*[2]
[1]Institute of Bioinformatics, University of Muenster, Muenster, Germany; [2]National Center for Biotechnology Information, National Institutes of Health, Bethesda, MD USAe-mail: wojmak@uni-muenster.de
* Corresponding author

Interspersed repetitive sequences are major components of eukaryotic genomes. They comprise over 50% of the mammalian genome. Because the specific function of these elements remains to be defined and because of their unusual 'behavior' in the genome, they are often quoted as a selfish or junk DNA. Our view of the entire phenomenon of repetitive elements has now to be revised in light of data on their biology and evolution, especially in the light of what we know about the retroposons. We argue that even if we cannot define the specific functions of these elements, we still can show that they are not useless pieces of the genomes. The repetitive elements interact with the whole genome and influence its evolution. They may serve as recombination hot spots or acquire specific cellular functions such as RNA transcription control or even become part of protein coding regions. Finally, they provide very efficient mechanism for genomic shuffling. As such, repetitive elements should be called genomic scrap yard rather than junk DNA.

Our recent focus is on Alu elements who are a particularly successful class of primate specific transposable elements. More than 1.2 million Alus are found in the human genome to which they contribute ~10% of its size.

It has been documented that when inserted into intronic regions, a few mutations are often enough to activate splice sites and determine inclusion of Alu fragments into mature mRNA molecules. Several studies revealed that Alu fragments are present in coding sequences of many genes, but a single case where a constitutively spliced Alu exon produces a functional protein has yet to be documented. Corroborated with Alu exons being alternatively spliced, this raises the question of whether Alu elements can contribute with functional domains to the host proteins. Interestingly, we found several examples when Alu cassettes potentially serve as templates for a signal peptide. Even more intriguing are cases of Alu cassettes promoting selenocysteine incorporation.

Here we take advantage of the recently sequenced macaque genome in order to investigate the pattern of selection acting on Alu exons. Even though publicly available SNP data are not useful for this task mostly because of the relative small size of the exons, the macaque genome presents divergence that allows for conducting the codon-based Fisher's exact test of selection. We did not find any exon subjected to significant positive or purifying selection, which together with other type of evidence (e.g. comparison with intergenic Alu sequences, protein homology modeling) suggests that Alu exons function as regulatory elements at mRNA level rather than adding new functionality to the host proteins.

INDEPENDENT COMPONENT ANALYSIS ALGORITHMS FOR MICROARRAY DATA ANALYSIS

Malutan R.*[1, 2], Vilda P.G.[1], Borda M.[2]
[1] Universidad Politecnica de Madrid, Boadilla del Monte, Spain
[2] Universitatea Technica Cluj Napoca, Cluj Napoca, Romania
[*] Corresponding author: e-mail: raul.malutan@com.utcluj.ro

Motivation and Aim: Microarray expression level scaling may suffer strong changes for a given gene or group of genes, which makes the small to be hidden in the large. Low expression levels should be as reliably estimated as high levels to monitor these changes and their possible cause and influence. Specific algorithms to estimate, detect and group genes acting together as metagenes are to be developed using independent component analysis and non negative matrix factorization as powerful tools having demonstrated their specific ability in treating these problems.

Methods and Algorithms: The amounts of hybridized material in an oligonucleotide microarray experiment for the perfect ($x_{i,k}^{p}$) and mismatch ($x_{i,k}^{m}$) probe pairs k corresponding to gene i given by the following relations: $x_{i,k}^{p} = \rho(s_{i,k}, x, y) p_t(s_{i,k} \mid z_{i,k}^{p})$ and $x_{i,k}^{m} = \rho(s_{i,k}, x, y) p_t(s_{i,k} \mid z_{i,k}^{m})$, will have under certain conditions that involved hybridization thermodynamics proportional probabilities. We will refer to these cases as reliably expressed cases. But in certain cases this assumption can not be checked, as there are some p-m pairs where strict proportionality does not match that of others within the same test probe. These cases may be referred to as unreliably expressed cases. The level of reliability for a probe set expression can be measured by the proportionality parameter of gene i, $\lambda_i = \frac{\left\|\mathbf{x}_i^m\right\| \cos\beta_i}{\left\|\mathbf{x}_i^p\right\|} = \frac{\left\langle \mathbf{x}_i^m, \mathbf{x}_i^p \right\rangle}{\left\|\mathbf{x}_i^p\right\|^2}$, while the orthogonality parameter may be measured by $\gamma_i = 1 - \cos^2\beta_i$. Using these two parameters, the microarray data can be study to determine unreliable gene expression samples. Unreliably expressed probe sets may be re-aligned using ICA algorithms, like FixedPoint ICA, AMUSE, JADE.

Results: For a single microarray containing information of PM-MM probe sets from 223391 gene positions, and having a large number of unreliable gene expression tests (γ_i>0.1), the re-estimated components show an improvement of co-linearity. When using FixedPoint ICA, an unreliably expressed probe with γ=0.143 was re-aligned, the new probes having orthogonality parameter of 0.0873.

Conclusion: The number of unreliable gene probe sets found in a particular microarray may be quite large, thus meaning that many probe tests may have been affected by corruption processes. These probe sets may be re-aligned by detecting their independent components by ICA, and re-estimating the PM-MM pairs from the independent components found. The results show that re-alignment improve the reliability of genes affected by underlying processes related with the independent components.

Availability: The software implementation of the algorithm is available on request from the authors, while the ICA tool used is available on http://www.bsp.brain.riken.jp/ICALAB.

AN EVOLUTIONARY AND COMPARATIVE GENOMICS BASED ACCOUNT OF Y-BOX PROTEINS IN EUKARYOTES

Mani A., Gupta D.K.*
Centre of Bioinformatics, University of Allahabad, Allahabad- 211002, India.
email:dwijenkumar@rediffmail.com
*Corresponding author

Motivation: Y-box proteins are a family of highly conserved nucleic acid binding proteins that are conserved from prokaryotes to human. They are defined as a group based on their ability to bind to the Y-box element. These proteins have been identified in various eukaryotic organisms. They are supposed to be involved in both transcriptional and translational control. Y-box proteins consist of three domains: the N-terminal domain, the cold shock domain and the C- terminal domain. Y-box proteins have been shown to interact with a number of cellular and viral proteins that are involved in various cellular processes. Expression of Y-box proteins has been correlated with response to cold stress, tumor progression, cell nuclear antigen proliferation in human lung cancer as well as with cell cycle progression. Little is understood about their evolution end genomic diversity among different taxa.

Aim: The study was performed by combining Bioinformatics and phylogenetic approaches in order to address first cross family evolution of y-box proteins among different eukaryotic organisms. A genomics based statistical approach was applied towards these functionally significant proteins to have an insight into conserved domains across and within the taxa of organisms.

Introduction: The Y-box proteins are the most evolutionarily conserved nucleic acid-binding proteins yet described, found in bacteria, plants and animals. The eukaryotic Y-box proteins were originally identified through their ability to interact with DNA containing a reverse CCAAT box, the Y-box sequence CTGATTGGCCAA. This sequence is found in a variety of promoter regions, including those of the MHC class II genes and genes encoding germ cell-specific functions and in these contexts the Y-box proteins are considered to act as regulators of transcription.

Methods and algorithms: In order to search Y-box protein family members we performed BLAST by using blastp program in the protein database at NCBI. *Mus musculus* y-box protein's gi|2745892|gb|AAB94768.1| amino acid sequence was selected as query. From the hits 17 sequences each from different species were selected for further studies. All the sequences were taken in FASTA format. The sequences were examined individually and aligned using CLUSTALW. Multiple sequence alignment, phylogenetic and molecular evolutionary analyses were conducted using MEGA version 4. For pair wise and multiple alignment gap open penalty was -7 and gap extension penalty was -1. BLOSUM weight matrix was selected for substitution scoring. Hydrophilic gap penalties were used to increase the chances of a gap within a run (5 or more residues) of hydrophilic amino acids; these are likely to be loop or random coil regions where gaps are more common. The aligned sequences were used to create phylogenetic tree. The evolutionary history was inferred using the Neighbour-Joining method. All the characters were given equal weights. The bootstrap consensus tree inferred from 10000 replicates was taken to represent the evolutionary history of the taxa analyzed. Branches corresponding to partitions reproduced in less than 50% bootstrap replicates were collapsed. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown next to the branches. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the poisson correction method and are in the units of the number of amino acid substitutions per site. All positions containing gaps and missing data were eliminated from the dataset (Complete deletion option). There were a total of 158 positions in the final dataset, out of which 87 were parsimony informative. Phylogenetic analyses were conducted in MEGA4.

*Results:*By statistical analysis of multiple aligned sequences it was observed that glycine, arginine, proline, alanine, glutamine and valine are the most frequently present amino acids with frequency percentage of 12.76, 12.03, 9.52, 8.21, 8.21 and 5.76 respectively. While within conserved sites glycine, valine, alanine, asparagine, glutamine, lysine and serine are the most frequently present amino acids with frequency percentage of 19.13, 14.87, 9.25, 8.50, 8.38, 6.37, 6.37 and 5.12 respectively. These Y- box proteins contain a cold shock domain which is 70 or 71 residues long and highly conserved. It was observed that within the cold shock domain valine, glycine, asparagines, lysine, glutamine and alanine are most frequently occurring amino acids with frequency percentage of 14.44, 11.37, 9.54, 9.38, 8.30 and 6.06 respectively. The multiple aligned sequence was found with No. of conserved sites=54, No. of parsimony informative sites= 230 and No. of singleton sites= 84.

A phylogenetic tree was constructed by using Neighbour –joining method.The tree shows different organisms on tree nodes branched on the basis of their Y-box proteins. *Schistosoma japonicum* makes a totally diverged branch from the main tree among 17 selected proteins. Node for Endopterygotans (*Chironomus tetanus, Bombyx mori, Drosophila melanogaster and aedes aegypti*) is supported by lower bootstrap values i.e. 68% while the node for vertebrates is supported by very high bootstrap support value i.e. 100%. Node for Teleosteians (*Carassius auratus*, *Danio rerio* and *Oryzus laetipus*) and mammals is supported by 99% bootstrap value. Node for mammals is supported by 100% bootstrap value, only exception is *Xenopus tropicalis* which is an amphibian.

GENES EXPRESSION EFFICIENCY ACCORDING TO ITS 5'-REGIONS AND CDS NUCLEOTIDE CONTENTS

Matushkin Yu.G.[1, 2]*, Likhoshvai V.A.[1, 2], Lashin S.A.[1, 2], Vishnevsky O.V.[1]
[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
[2] Novosibirsk State University, Novosibirsk, Russia
e-mail: mat@bionet.nsc.ru
* Corresponding author

Motivation and Aim: Correlation of gene expression with the degree of codon bias is known in many unicellular organisms. However, in a number of organisms such correlation is absent. Recently we have shown that consideration of inverted complementary repeats within open reading frames (ORFs) is necessary for proper estimation of translation efficiency of genes for some organisms.

Methods and Algorithms: The previously developed software suit of Internet-available programs providing estimation of potential expression level (index of elongation efficiency – IEE) for coding sequences in unicellular organisms' genomes was used for 612 organisms, which genomes were in GenBank for later 2007 year. IEE is calculated taking into account three factors: codon composition of coding sequence, its saturation of inverted repeats and free energy of potential hairpins.

Results: For each organism the optimal individual combination of these factors is found. The translation numerical characteristics of 612 sequenced unicellular organisms (563 bacteria, 49 archeans) are obtained. Taking into account the sample of 612 organisms the presence of five evolutionary strategies of translation optimization is confirmed. The significant difference of preferred strategies between bacteria and archeans is confirmed.

To improve the quantitative value of translation efficiency we used the structure of 5'-region of mRNA as additional information. The oligonucleotide potential (OP) was calculated for those regions (function of gene regulatory regions recognition, based on comparison of representation and characteristics of motifs distribution in investigated sequence and on gene regulatory regions sequences). The reliable correlation between OP and IEE is shown for *M.gallisepticum*. When approximated by cubic polynomial, the correlation coefficient was 0.24 ($p<0.001$). For *M.genitalium* the similar study has shown the correlation coefficient between OP and IEE to be 0.23 ($p<0.3*10^{-7}$) for all genes. However, 77 the worst OP sequences had the correlation coefficient with IEE if -0.19 ($p< 0.05$). Such effect can be explained with the species specificity of 5'-regions, *M. genitalium* is the smallest one of mycoplasms. To further improve of quantitative value of translation efficiency we have used the internet-available program UNAFold which calculates most probable secondary structures and their energy. In mRNA, the saturation of such structures differs in high- and low-expressing mRNA which were determined taking into account IEE values. The expression correlation of all genes of *H.pylori* with IEE growth with increase of gene length and that dependency is nonlinear. Maximal value $r =0.58$ is achieved when the length 2200 bp.

Conclusion and Availability: The proposed method allows to estimate numerically the coding DNA sequence expression efficiency and to optimize nucleotide composition of heterologous genes in all unicellular organisms. Web-version of the program is available: http://wwwmgs.bionet.nsc.ru/mgs/programs/eei-calculator/

Acknowledgements: Work was supported by RFBR (No. 06-04-49556), Project №13 of RAS Presidium program "Biosphere origin and evolution", Project №10.7 of RAS Presidium program №10.

MODELING THE EXPRESSION OF THE DROSOPHILA EVEN-SKIPPED (EVE) GENE DRIVED BY ITS PROXIMAL 1.7 KB UPSTREAM REGION

Matveeva A.D.*[1], Ionides J.M.C.[1], Reinitz J.[2], Samsonova M.G.[1]
[1] Department of Computational Biology, Center of Advanced Studies, St. Petersburg State Polytechnical University, 29 Polytechnicheskaya ul., St. Petersburg, 195251, Russia;
[2] Department of Applied Mathematics and Statistics, and Center for Developmental Genetics, Stony Brook University, Stony Brook, NY 11794-3600, USA.
e-mail: anya@odd.bio.sunysb.edu
* Corresponding author

Motivation and Aim: A central problem in modern molecular genetics is that of understanding how DNA regulatory sequences control gene expression. Metazoan regulatory regions are extremely complex and qualitatively different from those of prokaryotes. For example, the regulatory regions of genes controlling development in Drosophila are large and consist of groups of binding sites, called cis-regulatory modules (CRMs), each controlling some aspects of gene expression. The goal of our work is to understand the role of proximal 1.7 kb upstream regulatory sequence in the regulation of the Drosophila even-skipped (eve) gene in terms of binding sites.

Methods and Algorithms: To achieve this aim we have applied an approach which includes both mathematical modeling and experiment. We 1) quantitatively monitor gene expression at high resolution in space and time; 2) characterize transcription binding sites; 3) use new quantitative and predictive model of transcriptional readout of the proximal 1.7 kb of the Drosophila even-skipped gene to calculate the effect of a set of transcription factors bound to a large group of binding sites.

Results: As a starting point we consider the 34-site model that includes binding sites for repressors: Kruppel (Kr), Giant (Gt), Knirps (Kni) and Tailless (Tll) and activators: Bicoid (Bcd), Hunchback (Hb) and Caudal (Cad). This model was published in Janssens et al., 2006.

We propose that Hb can act as activator only if it is bound to the site located close to the binding site of Bcd, otherwise Hb acts as a repressor. To check this hypothesis we modify the 34-site model by considering that Hb bound to hb1 site acts as a repressor This modification improves the quality of fits. The rms of a new model was 8.66 and the features of experimental data were more accurately reproduced.

Next we consider the participation of Sloppy-paired 1 (Slp1) protein in the transcriptional regulation of *eve 2* stripe. Slp1 is expressed in a gap gene-like domain anterior to *eve* stripe 2. *In silico* addition of the Slp1 binding site (position from -1221 to -1205) to the 34-site model further decreases the rms value (rms=8.48) and improves the quality of patterns.

Conclusion: We introduce two new modifications in the previously published 34-site model of the p1.7*eve-lacZ* expression. These modifications significantly improve the quality of pattern prediction. Our results clarify the organization and regulation of the *eve* upstream region. The *in silico* modeling of transcription provides new insights into the mechanisms of gene regulation and allows to thoroughly analyze gene regulatory regions.

INTRON LANDSCAPE OF HUMAN GENOME

Maximov D.A., Babenko V.N.*
Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
e-mail: bob@bionet.nsc.ru
* Corresponding author

Motivation and Aim: Two facts are currently inferred in the course of eukaryotic genome investigations. The first finding is that the highly expressed genes in eukaryotes maintain short introns ([1],[2]). The second observation is that there is domain-wide regulation of gene expression in human, which comprises regions of ~80-90 genes per domain on average, exhibiting a particular level of integral expression ([3]). In this work we analyzed the features of various types of genes in regard to the issues mentioned above.

Methods and Algorithms: Intron length profiles were built along the human chromosomes and were overlapped with the abovementioned annotated expression domains. The profiles were calculated using high performance parallel computation mainframe itanium at **Institute of Computational Mathematics and Mathematical Geophysics SB RAS (www.sscc.ru).**

Results: We found that there are 2 distinct groups of genes in low expressed domains. One contains genes with longer introns. They are characterized with distinct Gene Ontology features. There is another class of moderately expressed genes which is often featured as cluster-structured gene loci with various activities, including immunospecific and cytotoxic ones. On the basis of intron length distribution we predicted novel domains of the particular expression level.

Conclusion: Our findings confirm the supposition that there are at least two modes of intron evolution.

Availability: The intron profiles are available at the institute web site: ftp://ftp.bionet.nsc.ru/pub/ip

References:

1. Petrov, D.A., Lozovskaya, E.R. & Hartl, D.L. High intrinsic rate of DNA loss in *Drosophila*. *Nature* **384**, 346–349 (1998).
2. Castillo-Davis C.I., Mekhedov S.L., Hartl D.L., Koonin E.V., Kondrashov F.A. Selection for short introns in highly expressed genes. *Nat Genet.* 2002;31(4):415-418.
3. Gierman H.J., Indemans M.H., Koster J., Goetze S., Seppen J., Geerts D., van Driel R., Versteeg R. Domain-wide regulation of gene expression in the human genome.
4. *Genome Res*. 2007;17(9):1286-1289.

3-D MODEL FOR AGROBACTERIAL T-DNA-BINDING VIRE2 PROTEIN

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

Motivation and Aim: Members of the genus *Agrobacterium* (family *Rhizobeaceae*) are natural soil-borne plant-root-system bacteria that can transfer a portion of their Ti-plasmid DNA (T-DNA) into host-plant nucleus under condition of virulence-gene activation. It is belived that VirE2 proteins form a membrane-spanning pore in lipid bilayer, and mediated ss-T-DNA-VirD2 protein complex delivering to the plant cell chromosomes [1]. Early, we revealed of a single, long-time jumps of membrane conductivity (or channel formation) during co-incubation flat dark membrane with VirE2 protein in a voltage-dependent manner [2]. How the VirE2-born pores are functioning is not known, however. The aim of this work was computer simulation of VirE2-protein 3-D models.

Methods and Algorithms: We used PHYRE 0.2. program and VirE2 protein primary sequence originated from NCBI {{http://www.ncbi.nlm.nih.gov}} for preparing the 3-D model for VirE2 protein. We used the Deep View Swiss-PdbViewer for visualization of prepared 3-D models.

Results: We prepared the 3-D model-segment for VirE2 started from 172 up to 529 amino acid sequence (64% of full-length VirE2 protein) using PHYRE program as a "band chains". According to this model a set of β-shits localized on the “left” side of VirE2 protein. A set of α-helixes localized mainly at the opposite parts of VirE2 protein. *BacterioRhodopsin* protein was applied as a prediction precision control for model prepared by PHYRE program. Using PHYRE program we can observed only two of 11 α-helixes of *BacterioRhodopsin* obtained by roentgeno-structure method. Nevertheless, we observed a high structure similarity for checked VirE2 fragment using Magic Fit and transfer methods. We observed a high structure similarity for checked VirE2 fragment.

Conclusion and Availability: Thus, using PHYRE program and primary amino acid sequences we prepared a high-precise 3-D model for VirE2 protein fragment started from 172 up to 529 amino acid (64% of full length protein).

References:

1. F. Dumas, M. Duckely, P. Pelczar, P. Van Gelder, B. Hohn (2001) An *Agrobacterium* VirE2 channel for transferred-DNA transport into plant cells, *Proc. Natl. Acad. Sci. USA* **98** : 485-490.
2. M.I. Chumakov, I.V. Volokhina (2005) Computer and experimental researching of Agrobacterial T-DNA binding VirE2 protein. In: *Proc. 2-nd Moscow Conf. on Computational Molecular Biology.* Moscow, Russia, July 18-21, 76-78.

SDPFOX: A TOOL TO PREDICT PROTEIN SPECIFICITY AND SPECIFICITY DETERMINANTS FROM MULTIPLE SEQUENCE ALIGNMENT

Mazin P.V.[1*], Mironov A.A.[1, 2], Rakhmaninova A.B.[1, 2], Gelfand M.S.[3], Kalinina O.V.[2, 3]
[1] Department of Bioengineering and Bioinformatics Moscow State University Moscow, Russia
[2] Institute for Information Transmission Problems RAS, Moscow, Russia
[3] EMBL-Heidelberg, Germany
e-mail: iaa.aka@gmail.com
[*] Corresponding author

Motivation and Aim: Protein functional annotation from their amino acid sequences is one of the big challenges in bioinformatics. Whereas prediction of general biochemical function is a well-studied issue, the specificity prediction field is currently in the phase of rapid expansion. The residues essential for protein function may be split into two types: those responsible for the general function of a protein and conserved within the whole family of homologous proteins; and those responsible for the specific recognition of the interacting molecule are presumably conserved among proteins with the same specificity but differ among proteins with different specificity. Let SDP (Specficity-Determining Position) be the alignment position with the latter conservation pattern. We assume that they play a role in protein specific interactions.

Methods and Algorithms: Recently, we developed SPDclust, a group of methods for indentification of SDPs in a multiple alignment of a protein family [1]. These methods enable analysis of protein families that lack information on specificity of their members. SDPclust includes four interconnected methods that allow to identify SDPs given an alignment and a grouping of proteins by specificity (SDPlight); assign proteins with unknown specificity to one of the pre-identified specificity groups (SDPprofile); group proteins by specificity given specificity assignment for a few family members (SDPgroup); and build a cluster tree of predicted protein specificities in absence of *a priori* knowledge on it (SDPtree).

Results and conclusions: We implemented SDPclust as a Web-server (SDPfox) and as a stand-alone console program (SDPproff) and compared its performance to a range of available methods that aim to predict protein specificity. Our benchmark on real and generated data shows that SDPclust performs equally or better than other methods. Moreover, its independence from phylogeny and overall sequence identity makes it applicable to a set of problem that none of the currently available tools is able to cope with.

Availability: SDPfox is available at the URL http://storage.bioinf.fbb.msu.ru/SDPfoxWeb/main.jsp , SDPproff can be downloaded from http://storage.bioinf.fbb.msu.ru/~mazin/SDPproff.jar

Acknowledgements: This work is supported by Howard Hughes Medical Institute (55001056), RFBR (07-04-91555). O.V.Kalinina is supported by the EMBO Long Term Fellowship (ALTF 119-2007).

References:

1. P.V. Mazin, et al. 2007 “SDPclust: a new tool for prediction protein specificity in MPA”. Moscow Conference on Computational Molecular Biology `07.

REDUCED LEVEL OF SYNONYMOUS SUBSTITUTION IN CPG CONTAINING CODONS SUGGESTS FUNCTIONAL ROLE OF INTRAGENIC AND 3' CPG ISLANDS IN HUMAN GENES

Medvedeva Ju.A.[1]*, Fridman M.V.[2], Oparina N.Ju.[3], Malko D.B.[4], Ermakova E.O.[5], Makeev V.Ju.[6]

[1] Institute of Genetics and Selection of Industrial Microorganisms, Russia
e-mail: ju.medvedeva@gmail.com
* Corresponding author
[2] Institute of Genetics and Selection of Industrial Microorganisms, Russia
e-mail: marina-free@mail.ru
[3] Engelhardt Institute of Molecular Biology, RAS, Russia, e-mail: oparina@gmail.com
[4] Institute of Genetics and Selection of Industrial Microorganisms, Russia
e-mail: carbonoid@mail.ru
[5] Institute for Information Transmission Problems, RAS, Russia,
e-mail: ermakova8@yandex.ru
[6] Institute of Genetics and Selection of Industrial Microorganisms, Russia
e-mail: makeev@genetika.ru

Motivation and Aim: CpG islands (CGIs) are defined as DNA segments longer than 200 bp, having over 50% of G+C content, and CpG frequency of at least 0.6 [1]. Most of the studies focused on CGIs considered only CGIs associated with 5' gene regions. In such CGIs many transcriptional factors binding sites have been found. The methylation status of 5' CGIs is believed to influence the transcription level of a corresponding gene.

Contrary to the widespread opinion only 40% of CGIs are located near TSS (5'CGIs). About 30% of CGIs are disposed in internal and 3' gene regions (intragenic and 3'CGIs). There are several examples that intragenic and 3'CGIs can perform important biological functions. Intragenic and 3'CGIs tend to overlap with coding exons more then with introns. Thus, the question arises if CGIs overlapping with protein-coding regions are in fact the result of selection at the protein level.

Methods and Algorithms: We compared selection at genome and protein levels by studying substitution rate between human and mouse orthologs in CpG sites belonging to 5', intragenic and 3' CGIs, using dn/ds test [2]. To this end we compared the substitution rate in exons overlapping and not overlapping with CGIs separately for different non-CpG containing codons (the background) and CpG containing codons.

Results and conclusions: 1. CGIs decrease the substitution rate in CpG pairs at synonymous sites approximately two-fold.

2. Effect of CGI does not depend on location within the gene: 5', intragenic and 3' CGIs protect CpG sites from extinction and probably play the same regulatory role in gene functioning.

References:

1. Gardiner-Garden, M. & Frommer, M. 1987. CpG islands in vertebrate genomes. *J. Mol.Biol*. 196, 261–282.
2. Ina, Y. 1995. New methods for estimating the numbers of synonymous and nonsynonymous substitutions. *J. Mol. Evol.* 40, 190–226.

PROTEIN FUNCTIONAL SITES PROJECTION ON EXON STRUCTURE OF GENE

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

Motivation and Aim: Even the first researches in the area of comparing of the gene and protein structure were made in the 70s, we do not have database that could join all of them. Only in the last few years two resources related to the GenBank and PDB were developed: XdomView[1] and SEDB. But the information about functional sites location is still very essential and incomplete, especially about their location on the gene structure. So we developed another method for mapping them on the gene structure and analyzed the result distribution.

Methods and Algorithms: At first, using the links from the PDBSite[2] and EnPDB we have found the entries in EMBL, than we used some filters to found the entries related to the similar sequences, species or alternative forms from these databases. Further, the EMBL entry coding sequence was artificially translated and aligned with protein sequence from PDB entry.

Results: To describe the results we used here next abbreviations: *FS* – functional site; *FA* – functional aminoacid(s) (aminoacid(s) entering into the site); *FE* – exon(s), containing FA.

The number of FA in the sample varies from 0 to 20 and the 97% of the sites belong to the groups containing up to 8 FA. In our sample we have seen that with the increase of exon number the number of FE also increase. When we analyzed the sites situated in only one exon, we saw that with the increase of the sequence length the length of the FE also increase. Then we found that the distribution of the number of FA per one FE is not casual. If we discern the discontinuity accurately we will see that the main part of the FA in sample locates in the neighboring exons that means the site could be considered as one functional unit that inherited from the gene structure.

Conclusion: In summary, we received the expected results that the functional units are clustered into one unit. For the most cases we suppose there are several units. And the number of FA does not depend on the functional exon length. Another point we discovered that functional aminoacids do not appear by chance. So we are going to continue our research with the hope to found more fine mechanisms of evolution of the functional units.

Work was supported in part by RFBR: 08-04-91313-IND_a, RAS presidium program “Molecular and cellular biology”, the grant “Systems biology: computer and experimental approaches.

References:

1. Vivek G. et al. (2003) XdomView: protein domain and exon position visualization. *Bioinformatics*, **19(1)**:159-60.
2. V.A. Ivanisenko et al. (2005) PDBSite: a database of the 3D structure of protein functional sites. Nucl. Acids Res. **33:** D183-D187.

ETHNOSPECIFIC DISTRIBUTION OF HAPLOTYPES FOR IVS2(+4) T/C, IVS4(-47) T/C, AND IVS5(-44) A/G SITES OF HFE GENE AND THEIR POTENTIAL SIGNIFICANCE IN SPLICING

Mikhailova S.V.*, Babenko V.N., Romashchenko A.G.
Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
e-mail: mikhail@bionet.nsc.ru
* Corresponding author

Aim: HFE protein is the main regulator of iron metabolism in human. Mutations in this gene are associated with hereditary hemochromatosis and some common diseases. It is proven that HFE interacts with different proteins. Various sets of HFE mRNA isoforms had been found at present. We tried to undermine the functional significance of widespread intronic polymorphisms of HFE gene. *Methods* We investigated frequencies for IVS2(+4)T/C, IVS4(-47) T/C, and IVS5(-44) A/G polymorphisms of HFE gene and their linkages with C282Y, H63D, and S65S mutations of the gene in Siberian populations. Context analysis was performed to reveal a potential influence of the intronic polymorphisms in splicing. *Results* Of the eight theoretically possible intronic haplotype variants of HFE gene, we identified only four, TTG, TTA, CTA, and CCA in various ethnic backgrounds in Siberia. We confirmed the linkage for the C282Y mutation with the TTG haplotype and H63D with CTA variant. Our novel finding was the association between the S65C mutation and CCA. The location of the three intronic polymorphisms at a critical distance (shorter than 50 bp) from the splice sites was found. The IVS2(+4) T/C polymorphism is located in intron near the donor site junction and it is supposed to able to affect the choice of the 5' splice site by U1 RNP during splicing. The IVS 4(-44) T/C polymorphism is located at position -5 of a possible donor splice site. Predicted pre mRNA variants from different Hfe alleles containing the retained intron 4 sequences would generate soluble form proteins of the HFE protein through the stop codons residing either in frame with the upstream exon at a distance of 5 nucleotides away from the exon 4 / intron4 boundary or out of frame at a distance 110 nucleotides from beginning of exon 4. An expression of HFE mRNA with the retained sequence of the intron 4 was observed earlier in hemochromatosis patients. The IVS5(-44) A/G polymorphism is located in proximity to the 3' splice site of the HFE intron 5 nearby the assumed branchpoint. *Conclusion* Analyzed intronic SNPs can affect HFE mRNA splicing isoforms pattern and change protein functions. It can explain the differences in phenotypic manifestations of HFE - associated diseases.

AUXIN REGULATION OF ITS OWN TRANSPORT DETERMINES THE ROOT TIP STRUCTURE IN PLANTS

Mironova V.V.[1*], Omelyanchuk N.A.[1], Fadeev S.I.[2, 4], Kogai V.V.[2], Yosiphon G.[3], Mjolsness E.[3], Likhoshvai V.A.[1, 4]

[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
[2] Institute of Mathematics, SB RAS, Novosibirsk, Russia
[3] Institute of Genomics and Bioinformatics, University of California, Irvine, USA
[4] Novosibirsk State University, Novosibirsk, Russia
e-mail: kviki@bionet.nsc.ru
* Corresponding author

Motivation and Aim: Arabidopsis root apical meristem (RAM) organizes spatial patterns of root cell differentiation by adding new cells with strict lineage relationships. RAM consists of quiescent center (QC) surrounded by stem cells (initials). Upward located vascular initials give rise to the actively divided cells of the provascular meristematic zone. Downward located columella initials supply to slowly sloughed columella cells. The QC cells do not divide. Monitoring of free auxin level in the root tip by reporter constructs showed a sharp concentration gradient with maximum of auxin response in the columella initial cells and lower activity in the columella [1]. Positions of both QC and auxin maximum relative to the most apical root cell maintains permanent in the course of root growth despite of continuous cell divisions in this region. Here we show by computer simulation that acropetal auxin transport facilitated by the PIN1 proteins suffices to produce and maintain auxin maximum in the root tip.

Methods and Algorithms: The 1D model for auxin stream along the central longitudinal axis of root was developed, which considered the following processes: auxin flux from the shoot, diffusion, dissipation and PIN1 regulated active transport of auxin [2]. Regulation of PIN1 activity was defined by Hill's functions. The cell-based model was made using the Dynamical Grammar, a multiscale modeling framework in which a system can be comprised of continuous and discrete elements [3].

Results: To explain the experimental data on cell divisions in the RAM we proposed that the putative substance Y regulates this processes altogether with auxin. Y is constitutively expressed in the QC, moves by diffusion and its degradation exhibits dose response to auxin. The curve for cell division rates over the increasing of Y concentration has the bell shape. The cell-based model maintains the stable auxin distribution with maximum at the 4th-5th cells from the root tip fitting to experimental data. Cell types along the longitudinal axis can be easily specified in the model by difference in auxin concentration and cell division rates.

Conclusion: The model reproduces maintaining of auxin maximum in the root tip and explains the positional control of root cell predetermination occurring along the root longitudinal axis. The model gives some predictions about the mechanisms of lateral root initiation, root regeneration and modifications in auxin distribution under different conditions.

References:

1. S. Sabatini et al. (1999) An auxin-dependent distal organizer of pattern and polarity in the *Arabidopsis* root, *Cell*, V. 99 P. 463–472.
2. V.A. Likhoshvai et al. (2007) Mathematical model of auxin distribution in the plant root, *Ontogenez,* V. 38 N. 6 P. 374–382.
3. E. Mjolsness and G. Yosiphon (2007) Stochastic process semantics for dynamical grammars**,** *Ann Math Artif Intell,* V. 47. P. 329-395.

PGNS-ROOT – A DATABASE ON EXPRESSION OF GENES IN PLANT ROOT DEVELOPMENT

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

Motivation and Aim: The domination of Arabidopsis thaliana, as the main object in plant developmental genetics is going to be followed and extended by studies of molecular backgrounds of developmental processes in other species. The current challenge is to explore whether the regulatory mechanisms that control plant development in model species have been conserved in non-model species. Natural variability can even enhance the efficiency of application of modern methods and techniques, for example, plant stem cell research is much more easily conducted on plants with large meristems (tomato, maize, cactus), tobacco is more prominent in experiments with plant tissue cultures in vitro. This wealth of multilayered, heterogeneous, and autonomous data demands integration in a systematized and classified form for further analysis. There are a number of databases both on genomes of single plant species (TAIR, Gramene, BarleyBase and others) and plant comparative genomics, for example PlantGDB. We are developing AGNS (Arabidopsis GeneNet Supplementary DataBase) on Arabidopsis gene activity and phenotypes as results of changes in gene activity in different genetic backgrounds [2]. Using AGNS as the basis we aim to elaborate a database, which will allow integrating the data on developmental genetics coming from different species. Due to quite similar anatomical structure of roots in different plant species starting from the database on gene activity in plant root development will make easy the establishment of the main database formats and tools.

Methods and Algorithms: The database input system was created as a Java application. The Berkeley DB XML with XQuery-based access is used for data processing.

Results: The PGNS (Plant GeneNetwork System) structure and format have been worked out to support both all AGNS preferences and new features extended the database facility. The PGNS contains the following sections. The Sequence DataBase accumulates data on mutations in plant genes with reference to related genomic databases. The Expression and Phenotype Databases describe gene expression patterns it root tissues detected by different methods and changes in root system phenotype, respectively, in wild type as well as in mutants, transgenic plants and under different conditions. The annotated papers are listed in the Reference Database. Vocabularies on anatomy and morphology of root system and root developmental stages are developed, where the common and species specific characters in root structure and development were denoted. Also the vocabularies on plant species, ecotypes and treatments have been worked out.

Conclusion: The PGNS has been created to integrate, systematize, and classify using the common frame the data on gene activities of different plants species.

Availability: http://wwwmgs2.bionet.nsc.ru/pgns

References:

1. N. Omelyanchuk et al. (2006) AGNS - a database on expression of arabidopsis genes, In: *Bioinformatics of Genome Regulation and Structure II.* Eds. N. Kolchanov, R. Hofestaedt, L. Milanesi, 433-442 (Springer Science+Business Media, Inc).

MODELING OF DRUG EFFECTS ON HEPATITIS C VIRUS REPLICATION IN AN HUH-7 CELL

***Mishchenko E.L.*[1*], *Bezmaternykh K.D.*[1, 2], *Likhoshvai V.A.*[1, 2], *Ivanisenko V.A.*[1, 2], *Kolchanov N.A.*[1, 2]**

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

Motivation and Aim: Application of mathematical modeling for investigation of drug effects on hepatitis C virus (HCV) replication.

Methods and Algorithms: Generalized chemical kinetic method [1] and the MGSModeller program [2] were used for model development and calculations. .

Results: We developed a mathematical model, which describes replication of subgenomic HCV replicon in an Huh-7 cell and incorporates the mechanisms of the action of inhibitors leading to blockage of the steps of HCV replicon translation, polyprotein processing, and replication of the HCV strand RNAs. The current mathematical model is more comprehensive than our previous [3] because it incorporates the effects of a greater number of potential drugs (inhibitors) alone or in combinations. The efficiency of the action of inhibitors on NS3 protease, NS5B polymerase, and EMCV IRES of HCV replicon was evaluated. The parameter estimates included the half-lives of HCV RNA, the active replicase complex, of this complex associated with the plus- or minus-strand RNA, polyprotein, depending on the inhibitor K_is value. The combination of a low affinity and a high affinity inhibitors at low concentrations resulted in a strong synergistic suppressive effect on HCV replicon level; in contrast, the inhibitor alone at the same concentrations was without effect. These inhibitor combinations taken advantage in designing of novel anti-HCV agents can be beneficial at relatively low concentrations. Combinations of agents would be helpful in combating HCV drug resistance and reducing the toxicity of potential drugs.

Conclusion: Our model's ability to demonstrate synergy in combinations of two inhibitors offers new prospects for the evaluation of the utility of potential anti-HCV agents. On the background of our results, we suggest that combinations of inhibitors with different mechanisms of action should be prioritized in evaluation of chronic hepatitis C antivirals.

References:

1. V.A. Likhoshvai et al. (2001). Generalized chemokinetic method for gene network simulation, *Mol. Biol.*, **35:**1072-1079.
2. F.V. Kazatcev et al. (2008). MGSmodeller'2008 – a computer system for reconstruction, calculation and analysis mathematical models of molecular genetic system . *This issue.*
3. E.L. Mihchenko et al. (2007). Mathematical model for suppression of subgenomic hepatitis C virus RNA replication in cell culture. *J. Bioinform. Comput. Biol*. **5:**593-609.

ESTIMATION OF MINIMAL DRUG TREATMENT DURATION FOR CLEARANCE OF AN HUH-7 CELL FROM HEPATITIS C VIRUS REPLICON BASED ON MATHEMATICAL MODELLING

***Mishchenko E.L.*[1*], *Bezmaternykh K.D.*[1, 2], *Likhoshvai V.A.*[1, 2], *Ivanisenko V.A.*[1, 2], *Kolchanov N.A.*[1, 2]**

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

Motivation and Aim: Long treatment of Huh-7 cells harbouring subgenomic hepatitis C virus (HCV) RNA replicon with inhibitors of HCV NS3 protease or NS5B polymerase resulted in HCV resistance and toxic effects on host cell. The aim of the study was to define minimal treatment duration for clearance infected cell from the HCV replicon using mathematical modeling.

Methods and Algorithms: Generalized chemical kinetic method [1] and the MGSModeller program [2] were used for model development and calculations.

Results: We developed a mathematical model, which describes replication of subgenomic HCV RNA replicon in an Huh-7 cell and incorporates the mechanisms of the action of inhibitors leading to blockage of the steps of HCV replicon translation, polyprotein processing, and replication of the HCV strand RNAs [3]. Calculations at 100 mcM of the inhibitors showed that for the high affinity inhibitors of NS3 protease (K_i = 1 pM - 0.63 nM) and NS5B polymerase (K_i = 1pM - 31 nM) treatment duration is brief (about 1 h) and weakly, if at all, dependent on K_i within the given range. Decrease in the affinity of inhibitors for targets - NS3 protease (K_i = 0.63 nM – 12 nM) and NS5B polymerase (K_i = 31 nM - 2mcM) - produced a fall in the minimum treatment duration from 1 h to 5 days within the analyzed ranges. A further relatively small decrease in affinity inhibitors of NS3 protease (K_i = 12 nM - 16 nM) and NS5B polymerase (K_i = 2 mcM - 2.5 mcM) caused a drastic increase in this parameter from 5 days to infinity within the examined ranges.

Conclusion: Our calculations showed that minimal treatment duration depends on the inhibitor constant K_i and allows, knowing the K_i inhibitor and its target, to estimate the minimal time for clearance of infected Huh-7 cells from subgenomic HCV RNA replicon.

References:

1. V.A. Likhoshvai et al. (2001). Generalized chemokinetic method for gene network simulation, *Mol. Biol.*, **35:**1072-1079.
2. F.V. Kazatcev et al. (2008). MGSmodeller'2008 – a computer system for reconstruction, calculation and analysis mathematical models of molecular genetic system . *This issue.*
3. E.L. Mihchenko et al. (2007). Mathematical model for suppression of subgenomic hepatitis C virus RNA replication in cell culture. *J. Bioinform. Comput. Biol*. **5:**593-609

A GRID ORIENTED EVOLUTION STRATEGY TO APPROACH THE PARAMETER ESTIMATION PROBLEM IN SYSTEMS BIOLOGY MODELS

Mosca E.[1], Merelli I.[1], Alfieri R.[1, 2], Milanesi L.[1*]
[1] Institute for Biomedical Technologies CNR, Via Fratelli Cervi 20090, Segrate, Milano, Italy
[2] Consorzio Interuniversitario per L'Elaborazione Automatica, Via R. Sanzio 20090, Segrate, Italy
e-mail: {ettore.mosca, ivan.merelli, roberta.alfieri, luciano.milanesi}@itb.cnr.it
[*] Corresponding author

Motivation and Aim: Nowadays a systems biology approach is essential to understand complex biological processes, such as cell cycle regulation [1]. This approach relies on mathematical models used to describe biological systems and make useful predictions. Due to the lack of experimental measurements, experimental errors and biological variability, the value of many parameters of the models is yet unknown or uncertain [2]. Although the computational approach can be successfully applied to identify these constants, the computational cost is very high for complex models. Here we present a grid oriented approach to compute parameter estimation of Ordinary Differential Equations (ODEs) models using an evolution strategy algorithm.

Methods and Algorithms: The parameter estimation is stated as the global optimization (GO) problem to find the parameter values which determine the best model fitting to experimental data. To solve the GO problem a well established method relies on using an Evolution Strategy. In particular, we chose the Stochastic Ranking Evolution Strategy algorithm (SRES) [3], which has shown good performance when applied to analysis of biochemical pathways [4]. Although this algorithm is very efficient, it can take long time to reach the global minimum of the optimization problem. An interesting approach to reduce the execution time relies in running different computations of the algorithm simultaneously, crossing periodically the best results among the processes, to speed up the convergence to the optimal solution. In this work we develop an environment to distribute each run of the evolution algorithm on a different grid working node. The key feature of the implementation is a relational database that allows the user to swap the individuals (i.e. solutions of the GO) among the working nodes during the computations.

Results and Conclusions: We have developed an automated system to manage the parameter estimation of ODE models oriented to grid computing. The system is made of two components: the application for parameter estimation and the application which manages the distribution over the grid working nodes. Preliminary results indicate that this approach can successfully lead to the parameter estimation of more complex ODE models, overcoming both the computational load and the difficulty of the GO problem.

Availability: free for academic use.

References:

1. R. Alfieri, I. Merelli, E. Mosca, L. Milanesi (2008) The cell cycle DB: a systems biology approach to cell cycle analysis, *Nucleic Acids Res.* 36(Database issue): D641–D645.
2. W. Liebermeister, E. Klipp (2005) Biochemical networks with uncertain parameters, *Systems Biology, IEE Proceedings*, 152: 97–107
3. T.P. Runarsson, X. Yao (2000) Stochastic Ranking for Constrained Evolutionary Optimization. *IEEE Transactions on Evolutionary Computation*, **4**: 284-294.
4. C.G. Moles, P. Mendes, J.R. Banga (2003) Parameter Estimation in Biochemical Pathways: A Comparison of Global Optimization Methods. *Genome Research*, **13**: 2467-2474

ASF1 FACILITATES H3K4 DEMETHYLATION BY A NOVEL COMPLEX SPEL IN GENE REPRESSION

Moshkin Y.[1]*, Kan T.W.[1], Goodfellow H.[2], Secombe J.[3], Eisenman R.N.[3], Bray S.J.[2], Verrijzer C.P.[1]

[1] Erasmus University Medical Center, Rotterdam, the Netherlands;
[2] University of Cambridge, Cambridge, UK;
[3] Fred Hutchinson Cancer Research Center, Seattle, USA
e-mail: i.mochkine@erasmusmc.nl
* Corresponding author

Histone chaperones regulate a diverse set of chromatin remodeling activities occurring DNA replication/repair and transcription activation/repression. Asf1 is likely to be the major histone H3/H4 chaperone and it is involved in both replication-dependent and independent chromatin assembly. However, Asf1 also plays a role in transcription through facilitating the eviction of histones from actively transcribed genes and as we have recently shown, it is also required for the repression of Notch target genes. These observations pose the major question of how Asf1 can regulate such a diverse set of chromatin remodelling processes given its relatively simple biochemical activity, the binding and releasing of histones. To address this question, we set up an unbiased screen for Asf1 interacting partners by purification of Asf1 complexes from Drosophila nuclear extract combined with the high throughput mass-spectrometry analysis. This approach turned to be effective, as besides previously known Asf1 interactors such as CAF-1 and HirA, we identified a set of novel Asf1-associated proteins. These include transcription regulators such as Sin3A, a PHD-finger protein PF1, a protein homologous to the human oncogene EMSY and a histone demethylase of JARID1 family – LID. Further biochemical analysis revealed that Sin3A, PF1, EMSY and LID form a novel complex SPEL (Sin3A-PF1-EMSY-LID). A fraction of the SPEL associates with Asf1 and we refer to this Asf1-containing complex as A-SPEL. A-SPEL complex contains histone H3K4 demethylase activity suggesting its function in gene repression. Therefore, we wondered whether Asf1 could facilitate histone demethylation by the SPEL complex in repression of the Notch target genes. Indeed, it appeared that subunits of A-SPEL are required for the repression of the Notch-inducible genes and depletion of cells for Asf1 and other SPEL subunits by RNAi leads to the increase of H3K4 methylation levels on enhancers of the Notch target genes. Together, our results suggest that Asf1 facilitates demethylation activity by the SPEL enzymatic subunit LID and provide a novel link between histone chaperones and histone modifying enzymes in control of gene silencing.

STATISTICAL ESTIMATION OF ERRORS IN GENE EXPRESSION DATA ARISING IN COURSE OF CONFOCAL SCANNING

Myasnikova E.M.*, Surkova S.Yu., Samsonova M.G.
St.Petersburg State Polytechnic University, St.Petersburg, 195251, Russia
e-mail: myasnikova@spbcas.ru
* Corresponding author

Motivation and Aim: The confocal scanning microscopy in conjunction with fluorescent labeling is a powerful tool for acquisition of accurate and standardized quantitative data at a resolution of a single cell. However the use of this technique for systems biology is limited due to possible experimental errors. The aim of this work is to estimate the most essential errors which arise in the course of fluorescence quantification and to set the range of experimental conditions allowing to obtain the sufficiently accurate data. We have developed statistical methods for estimation and correction of these errors using images of gene expression in embryos of *Drosophila melanogaster* as an example application [1].

Methods and Algorithms: (1) In photon-limited confocal imaging the major source of errors is Poisson noise due to the discrete nature of photon detection [2]. The common way to reduce this noise is averaging of multiple frames. We propose a method for the extraction, estimation and removal of the residual photon noise from averaged images and show how these errors are specified by the adjustment of microscope detector. (2) Averaging of images may also cause a loss of information due to clipping of single frames. A method based on censoring technique is used to estimate and correct errors stipulated by averaging of saturated pixels both at high and low intensities. (3) One more source of errors lies in the image segmentation procedure which results in the underestimation of expression levels when is applied to blurred confocal images. We propose a modification of Richardson-Lucy (RL) deconvolution method [3] which allows to estimate and correct these errors and makes the quantification procedure less sensitive to the accuracy of image segmentation.

Results and Conclusions: It has been shown that the attempts to improve an image quality by setting too high values of the microscope parameters yield the higher noise and more severe distortions of the data. The presented statistical methods allow to estimate and correct the data errors and set the admissible range of adjustable parameters.

An important application of the current work is the possibility to ensure the high accuracy of a large amount of quantitative gene expression data accumulated in the previous study (http://urchin.spbcas.ru/flyex/) as well as to give recommendations how to achieve the required accuracy of the data in the context of the further work. It has been proved that the choice of proper experimental conditions provide the high data quality sufficient to use the dataset to study mechanisms of pattern formation, to infer regulatory interactions in the genetic network and to develop new mathematical models.

Funding: This work is supported by NIH grant RR07801, GAP award RBO-1286 and NWO-RFBR project 047.011.2004.013.

References:

1. E.Myasnikova et al. (2001) Registration of the expression patterns of *Drosophila* segmentation genes by two independent methods, *Bioinformatics*, **17:** 3-12.
2. J.Pawley (2006) Fundamental limits in confocal microscopy, In *Handbook of biological confocal microscopy,* J.Pawley (Eds), 20-41 (Springer-Verlag New York Inc.).
3. G.M.P.van Kempen et al. (1997) A quantitative comparison of image restoration methods for confocal microscopy, *Journal of Microscopy*, **185:** 354—365.

ON THE OPTIMALITY OF THE GENETIC CODE: THE ROLE OF NONSENSE CODONS

Naumenko S.A.
Keldysh Institute of Applied Mathematics RAS, Moscow, Russia
e-mail: sergey_clark@yahoo.co.uk

Motivation: A role of nonsense codons in genetic code optimality is not underlined clearly at present. The aim of the research is to determine the role of nonsense codons in the process of minimization of logical translation errors.

Methods and Algorithms: We use a simplest model of the genetic code named the genetic code markup, which is the division of codon set into sense and nonsense parts. There is no difference between sense codons in the markup: all of them are equal. We use two optimization parameters: the probability of nonsense codon appearance in the result of frameshift mutation and the number of sense codons for which nonsense point mutations are possible.

Results: The standard genetic code markup terminates the translation process in the case of frameshit mutation with the highest probability in comparison with other markups. At the same time the standard genetic code markup has the minimal number of sense codons for which nonsense point mutations are possible. There are only 5,8% of optimal markups including the standard markup. This fact supports the adaptive genetic code hypothesis. Probably the optimal markup fixation on the early stages of evolution could provide the translation stability in the case of primitive molecular apparatus.

THE GH31 FAMILY OF GLYCOSIDE HYDROLASES: SUBFAMILY STRUCTURE AND EVOLUTIONARY CONNECTIONS

Naumoff D.G.
State Institute for Genetics and Selection of Industrial Microorganisms, Moscow, Russia
e-mail: daniil_naumoff@yahoo.com

Motivation and Aim: GH31 is a family of glycosidase catalytic domains, having the TIM-barrel type of the 3D structure. It is represented by proteins widely distributed among living organisms and possessing at least six types of enzymatic activities. Currently, sequences of about one thousand members of this family are deposited in web-databases. However, subfamily structure of the GH31 family and its relationship with other families are still unclear.

Methods and Algorithms: Protein sequences were retrieved from the NCBI database. The phylogenetic trees were built using programs of PHYLIP package. Precise boundaries of GH31 domains were determined according to PDB database (2G3M, 1WE5, and 2QLY) or by homology. In order to group GH31 domains, a representative of the family was used as a query for PSI-BLAST search. All sequences obtained during the first iteration with *E*-value less than 10^{-50} were considered as belonging to the same group as the query sequence. Then a new query for PSI-BLAST was randomly selected among the domains, which were still out of grouping. The procedure was subsequently repeated until all GH31 domains were included into at least one group. Note that protein fragments were not used as a query. At the next stage the groups of GH31 domains, containing at least one common member, were combined. Interfamily relationships were established by PSI-BLAST searches, using several representatives of the GH31 family and 0.005 as a threshold *E*-value for including a sequence in the next iteration. The number of iterations needed to reach a family member was considered as a degree of sequence similarity between GH31 and the corresponding family.

Results and Conclusion: On the basis of PSI-BLAST searches 59 groups of closely related domains were distinguished inside the GH31 family. Joining up of the overlapping groups of domains allowed reducing their number to 34, which we propose to consider as 34 subfamilies, forming the GH31 family. The biggest subfamily combines 16 groups, but the others contain only 1-4 groups. Phylogenetic analysis showed that each subfamily forms a separate cluster on the GH31 tree with exception of the biggest subfamily. The latter contains more than a half of the known GH31 domains and its cluster on the tree includes several (depending on the method of tree building) other subfamilies. It suggests that proteins of this subfamily probably were evolutionary ancestor for all other members of the GH31 family. Searches of the NCBI database by PSI-BLAST, using sequences of three GH31 domains with known 3D structure as a query, during the first eleven iterations revealed 5703 nonidentical protein sequences in total (i.e. about 0.1% of all known sequences), belonging to nine protein families. The results allowed to conclude that GH13, GH27, GH36, and COG1649 families are the closest to the GH31 family, while GH66, GH101, COG1306, and COG3868 are only distantly related to it. Relationship of GH31 with GH101, COG1306, and COG3868 families has been found for the first time.

I am thankful to Marco Carreras for help with automation of PSI-BLAST output analysis and to the Russian Foundation for Basic Research for a financial support (grant 06-04-49079-a).

GAG RELATED GENE HAS A CONSERVATIVE FUNCTION IN DROSOPHILA GENOME

Nefedova L.N.*, Kim A.I.
Moscow State University, Moscow, Russia
e-mail: lidia_nefedova@mail.ru
* Corresponding author

Motivation and Aim: For a long time the question of retroviruses' origin remained open. Coming from structural likeness and homology of separate genetic structural components of retrotransposons and retroviruses, it is possible to suppose that an evolutional process can go in two directions: both by transformation of retrotransposons to the retroviruses and by forming of retrotransposons on the base of retroviruses [1, 2]. If the process of assembling of retrotransposons and retroviruses went in the host genome, in it, probably, there must be homologues of their open reading frames (ORFs): *gag*, *pol* and *env*. But it's difficult to prove if these homologues are precursors or domesticated genes of retrotransposons. *Drosophila* genomes contain the homologue of the *env* gene (named *Iris*) that assumed to be acquired at baculoviruses [3]. We carried out the search and analysis of the *gag* gene homologues in *Drosophila* genomes.

Methods: of bioinformatics (search of homologues, multiple alignments, amino acid sequences analysis) were used as well as PCR- and RT-PCR-analysis.

Results: Unique homologue of the *gag* gene (*CG4680*), which encodes product with an unknown function, has been detected in the *D.melanogaster* genome. We named this gene *Grp* (*Gag related protein*). Homologues of the *D.melanogaster Grp* gene present at all investigated species of *Drosophila* with the different degree of phylogenetical relation. But at least sequenced genomes of the other insects don't contain homologues of this gene. So the age of the *Grp* gene is about 40 million years.

Surprising is that unlike the *Iris* gene, which, presumably, is under the action of positive selection, the *Grp* gene differs high conservatism of the sequence at all investigated species and is under purifying selection. It was shown the *Grp* gene is expressed on the transcription level. Very low level of transcription observed during the embryo stage of development and a high level - in adult flies.

Conclusion: Data obtained is the evidence of the conservative function of the *Grp* gene for *Drosophila*. Origin of the *gag* gene in *Drosophila* genome is discussed.

References:

1. L.N.Nefedova, A.I.Kim (2007) Evolution from retrotransposons to retroviruses: origin of the *env* gene, *Zh. Obshch. Biol*., **68**(6):459-467.
2. L.N.Nefedova, A.I.Kim (2007) Evolution of errantiviruses of *Drosophila melanogaster*. Strategy 2: from retroviruses to retrotransposons, *Genetika*, **43**(10):1388-1395.
3. *H.S.Malik, S.Henikoff (*2005) Positive selection of *Iris*, a retroviral envelope-derived host gene in *Drosophila melanogaster*, *PLoS Genet.,* **1**(4):e44.

THE NEW TOOL FOR OLIGONUCLEOTIDE DESIGN FOR VIRUS GENOTYPING USING MULTIPLE ALIGNMENT

Neverov A.D.*[1], Orlov S.G.[1], Mironov A.A.[2], Chulanov V.P.[1]
[1] Central Research Institute of Epidemiology, Moscow, Russia.
[2] Lomonosov Moscow State University, Bioinformatics and Bioengineering Department, Moscow, Russia.
e-mail: neva_2000@mail.ru
* Corresponding author

Motivation and Aim: Determination of microorganism genotypes isolated from clinical or environmental specimens often has epidemiological or clinical significance. The gold standard for genotyping is sequencing followed by phylogenetic analysis. In many cases it is reasonable to design genotyping assay based on RealTime PCR or Hybridisation technologies. Both methods use oligonucleotides selectively discriminating genotypes. We elaborate a computer program to ease a time consuming theoretical analysis of thermodynamic properties of olygonucleotides.

Methods and Algorithms: Multiple alignment is a source of information about sequence divergence/similarity. For variable organisms like RNA viruses we have to analyze large number of sequence to ensure presentation all possible alleles in each genome locus. To prepare data for analysis we sort sequences in alignment according their similarity and split alignment on groups corresponded to genotypes. The optimal probe for genotyping should have mach more stable duplexes on isolates of desired genotype compared to others. The probe should situated in a region conservative in the desired genotype and occupy positions where nucleotides not evenly distributed across genotypes – genotyping positions. The program looks for positions in alignment where nucleotide distribution is significantly different from stochastic. These positions have a valuable portion of phylogenetic signal about genotype subdivision. Searching of genotyping positions based on either mutual information or Goodman and Kruskal lambda-max statistics. In a practical case we are often interested to find positions where nucleotide distribution is significantly different in two groups of genotypes, say in genotype A and in all other genotypes of Hepatitis B virus. Rule for discrimination is represented in form of two lists of genotypes. We used a lambda-max statistics to check rule satisfiability. Note what the lambda-max statistics takes value from the range [0,1]. Nucleotide distribution across genotypes for each alignment position may be represented by contingency table. Table rows correspond to nucleotide types, columns – to genotypes. The zero value of lambda-max indicates absence of association between rows and columns, the value 1 corresponds to diagonal table. Algorithm constructs a binary tree joining columns (genotypes) at each step least distinguishable from each other (with minimal value of lambda-max). At each step table likelihood is calculated using lambda-max as a measure of table goodness multiplied by a penalty for column joining. The table corresponded to a step where the likelihood takes maximum value used for calculation of the satisfiability of discriminating rule. The rule is true (there is a significant difference in nucleotide distribution) if there isn't a pair of genotypes one from the 1st list of the rule and another from the 2nd joined together in the same supergenotype.

Results: The program has convenient graphical interface to show aligned sequences and nucleotide distributions in alignment positions in each genotype. Thermodynamics parameters (Tm, dG, dH) are calculated for each of possible heteroduplex at the site of probe annealing. User may explore the melting temperatures of the probe on sequences belong to different genotypes.

A MODEL STUDY OF THE ROLE OF PROTEINS CLV1, CLV2, CLV3, AND WUS IN REGULATION OF THE STRUCTURE OF THE SHOOT APICAL MERISTEM

Nikolaev S.V.[a*], Penenko A.V.[b], Lavreha V.V.[a], Smal P.A.[c], Mjolsness E.D.[d], Kolchanov N.A.[a]

[a] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia,
[b] Institute of Computational Mathematics and Mathematical Geophysics, SB RAS, Novosibirsk, Russia,
[c] Novosibirsk State University, Novosibirsk, Russia,
[d] Institute of Genomics and Bioinformatics, University of California, Irvine, CA 92607 USA
e-mail: nikolaev@bionet.nsc.ru
[*] Corresponding author

Motivation and Aim: There is a hypothesis that products of *CLV* genes inhibit *WUS* gene expression, and *WUS* gene product activates expression of *CLV*. Further more this interaction is supposed to be the basis for shoot apical meristem maintenance. To test this hypothesis a mathematical model of mechanism of these genes interaction was proposed and studied.

Methods and Algorithms: The model consists of identical molecular-genetic regulation circuits "working" in each cell. The circuits interact by gene diffusible products, called regulators (i.e. this is a reaction-diffusion model). The shoot apical meristem is represented by a semi-elliptic domain filled with polygonal cells (vertical section of the top of meristem). Mathematical model was defined by a system of ordinary differential equations. Cell geometry was used to calculate diffusion resistances for regulatory molecules transportation between the cells.

Results: A hypothetical substance Y and additional interactions had been introduced to complete mathematical model of *CLV-WUS* - based mechanism for shoot apical meristem maintenance. A set of parameters had been obtained that provide steady state solution for the model (spatial distribution of molecule concentrations). This solution is in accordance with experimentally observed concentrations of CLV3 and WUS proteins in shoot apical meristem.

Conclusion: The following regulation circuits are postulated in proposed model: (1) Y activates expression of *WUS* gene, (2) CLV inhibits *WUS* expression, and (3) WUS activates *Y* expression, Y, in turn, activates *CLV* expression. The numerical runs of the model demonstrate that postulated interactions between *Y*, *CLV* and *WUS* embedded in the reaction-diffusion mechanism are able to provide an appropriate "mark up" for the zones of stem cells and their niche in shoot apical meristem.

Acknowledgements: This work was supported by the RFBR grant: 08-04-01214-a "Mathematical modeling and analysis of structure homeostasis mechanism of the stem cell niche in Arabidopsis thaliana shoot apical meristem"

DE NOVO PREDICTION OF ALTERNATIVE SPLICING EVENTS WITHIN SH3 DOMAINS OF PROTEINS FROM DIFFERENT ORIGINS

Nikolaienko O.V.*, Dergay M.V., Dergay O.V., Morderer D.Y., Tsyba L.O., Skrypkina I.Y., Rynditch A.V.
Institute of Molecular Biology and Genetics NAS of Ukraine, Kyiv, Ukraine
e-mail: rynditch@imbg.org.ua
* Corresponding author

Motivation and aim: SH3 domains that are often found in proteins implicated in cell signaling pathways, cytoskeletal organization and membrane traffic are important for the assembly of multiprotein complexes. SH3 domains are highly adaptable. Phage library screens demonstrated that substitution of two or three residues within an SH3 domain is sufficient to alter its specificity. It is considered that RT and n-Src loops of SH3 domains that are highly variable in sequence and flexible in structure play important roles in modulating the specificity of SH3 domains.

Previously we have shown the presence of two isoforms of the SH3A domain of endocytic protein ITSN1 generated by neuron-specific alternative splicing of exon 20 [1]. Inclusion of exon 20 extends the n-Src loop of the ITSN1 SH3A domain and changes it`s ligand binding specificity. A similar brain-specific splicing event within an SH3 domain has been described for the Src protein, in which 18 nucleotides encoding six amino acids are inserted into the n-Src loop of the SH3 domain.

Methods and algorithms: In order to find similar alternative splicing events within other SH3 domains we developed the "pipe" consisting of several Perl scripts. On the first stage for consequent analysis genes, which meet the following requirements, were selected:

1) NCBI RefSeq data for corresponding mRNA and protein is available;
2) protein contains one or more SH3 domains encoded by at least 2 exons;

On the second stage introns between above-mentioned exons were scanned for constitutive splicing sites (GT and AG) and possible alternative exons meeting following criteria were selected:

1) exon size is from 9 bp to 27 bp and divisible by 3;
2) exon should not contain in-frame stop codon;

On the third stage for every predicted exon following parameters were evaluated:

1) "stringency" of 3` splice site (similarity to conservative 3` splice site);
2) "stringency" of 5` splice site (similarity to conservative 5` splice site);
3) conservation of corresponding region between genes from different origins (based on NCBI HomoloGene)

Results: Only SH3 domains of aforementioned ITSN1 and c-Src had high values of computed parameters. For complementary analysis few other top-scored genes (Yes1, Fyn1, Sorbs3, Scap1, Scap2, ArhGEF5) were selected. However, RT-PCR analysis with specific primers hasn`t revealed such alternative splicing events. Nevertheless, negative result can be explained by incomplete dataset (see stage 1 restrictions) and limited HomoloGene data available.

Availability: All data and scripts are available on request from the authors.

References:

1. Tsyba, L., Skrypkina, I., Rynditch, A., Nikolaienko, O., Ferenets, G., Fortna, A. and Gardiner, K. (2004) Alternative splicing of mammalian Intersectin 1: domain associations and tissue specificities. Genomics 84, 106-113

PROF_PAT, THE UPDATED DATABASE OF PROTEIN FAMILY PATTERNS – AN EFFECTIVE TOOL FOR GENOME ANNOTATION

Nizolenko L.Ph.[1]*, Bachinsky A.G.[1], Yarygin A.A.[1], Grigorovich D.A.[2]
*Corresponding author
[1]SRC VB "Vector", Koltsovo, Novosibirsk, Russia
e-mail: nizolenko@vector.nsc.ru
[2] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
e-mail: odip@bionet.nsc.ru

Motivation and Aim: **Pro**tein **f**amily **pat**terns bank Prof_Pat is a collection of the patterns of groups of related proteins, characterizing the position intervals conservative in aligned proteins, and flexible fast search program. It is one of the numerous "secondary" banks, in which the information on the whole groups of the related proteins, most typical and frequently unique features of this group is concentrated. Prof_Pat was earlier showed to have as good completeness and variety of included proteins as the best world-known "secondary" banks. At the same time, its specificity and sensitivity is higher than those of other banks, and its search speed was 3-10 times higher. [1]. In addition, Prof_Pat can examine large groups of protein sequences rather than just a few of them and so it appear to be an effective tool for genome annotation.

Methods and Algorithms: Protein family patterns, the bank of these patterns and flexible fast search program were created using original technology [2]. The version of Prof_Pat 1.19, constructed on the basis of the UniProt 8th release, contains patterns of 162161 groups of related proteins including more than 1332000 amino acid sequences.

Results: Prof_Pat is an updated tool for a prediction of function and distant similarity of proteins. The efficacy of Prof_Pat was demonstrated in analysis of amino acid sequences, translated from complete genomes of microorganisms:

- From 4105 sequences of *Bacillus subtilis* 99.88% were recognised by Prof_Pat. 22 sequences were annotated with high significance level for the first time. This protein's functions were not predicted up to now.
- From 3924 sequences of *Mycobacterium tuberculosis*, 99.95% were recognised by Prof_Pat. 44 sequences were annotated with high significance level for the first time [3].
- From 4767 sequences of *Salmonella typhi* + plasmides 99.94% were recognised by Prof_Pat. 16 sequences were annotated with high significance level for the first time.

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

Conclusion: Bank Prof_Pat, an updated, developing and improved tool for a prediction of function and relationships of proteins allows to get new information for assumption of structural and functional similarity for distinct proteins as well as for large groups of amino acid sequences.

References:

1. L. Ph.Nizolenko et al. (2003) Database of patterns PROF_PAT for detecting local similarities In Silico Biology, **3**, 205-213.
2. A.G. Bachinsky et al. (2000). PROF_PAT 1.3: updated database of patterns used to detect local similarities. *Bioinformatics*, **16**, 358-366.
3. L. Ph. Nizolenko et al. (2005a) Study of the amino acid sequences of open reading frames of the complete genome of Mycobacterium tuberculosis using the protein family pattern bank Prof_Pat. *Biofizika*. **50,** 986-992. (Russian).

INVESTIGATION OF THE AMINO ACID SEQUENCES OF HUMAN INFLUENZA VIRUS H5N1 WITH PROTEIN FAMILY PATTERNS BANK PROF_PAT

Nizolenko L.Ph.[1]*, Bachinsky A.G.[1]
*Corresponding author
[1]SRC VB "Vector", Koltsovo, Novosibirsk, Russia.
e-mail: nizolenko@vector.nsc.ru

Motivation and Aim: The purpose of the investigation was revealing similarity of an inflenza virus proteins with proteins of eucariotes and *Homo sapiens* in particular, as such similarity can negatively affect safety of vaccines against this virus.

Methods and Algorithms: Sequences of human influenza virus H5N1, circulated in different regions of the Earth in 2006 are available in the Internet from the NCBI (ftp://ftp.ncbi.nih.gov/genomes/INFLUENZA/).

Protein family patterns, the bank of these patterns and flexible fast search program were created using original technology [1]. Prof_Pat release 1.19, constructed on the basis of the UniProt 8th release, contains patterns of 162161 groups of related proteins including more than 1332000 amino acid sequences.

In analyzing we used a 250PAM similarity matrix, with 80% degree of similarity as the lower boundary. Similarity was considered positive if not less than two ordered motives of the pattern in length on 10 amino acids identified corresponding sequence.

Results: Similarity of neuraminidase of CDC623 strain (Indonesia) with beta carotene oxygenase family, including human and other vertrebrates sequences was found. Polymerase PA of many Indonesian strains contains the fragments similar with dynein heavy chain of vertebrates, and polymerase PB1 of Indonesian strain CDC739 – with mammalian bone marrow stromal antigen 2.

In addition, the similarity of influenza virus proteins with proteins of different eucariotes (such as Saccharomycetes, Plasmodium and plant mitochondia), bacteria and viruses was reveald.

Still time significant variability of influenza virus proteins is noted, and it concerns not only hemagglutenin and neuramimidase, but also polimerases and nucleocapsid protein.

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

Conclusion: Results the investigation can be used in practical pharmacology with the purpose of improvement of quality anti-influenza vaccines and reduction possible postvaccinal accidents.

References:

1. Bachinsky A.G., et al. PROF_PAT 1.3: Updated database of patterns used to detect local similarities. Bioinformatics, 2000, 16, 4, 358-366.

HORIZONTAL TRANSMISSION OF NON-LTR RETROTRANSPOSONS: ARTEFACT OR RARE EVENT?

Novikova O.*, Blinov A.
Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
e-mail: novikova@bionet.nsc.ru
* Corresponding author

Motivation and Aim: Horizontal transfer (HT) can be defined as the process by which genes can move between reproductively isolated species. It is not surprising that many examples of HT of mobile elements have been identified in eukaryotes. Mobile elements have the capacity to insert themselves into the chromosomes of possible vectors and, subsequently, into host chromosomes. Nevertheless, non-LTR retrotransposons are believed to be inherited exclusively vertically. Extensive phylogenetic and comparative studies dismissed majority of putative HT reports for non-LTR retrotransposons. At the same time, strong evidence of HT that cannot be neglected were provided for Bov-B retroelements which have been transferred from the ancestral snake lineage (Boidae) to the ancestor of ruminant mammals [1] and for CR1B elements which have been transmitted between Bombycidae moths and Lycaenidae butterflies less than 10 MYA [2]. Thus, one should not exclude a possibility of occasional HT events in non-LTR retrotransposons. In present report, one more case of non-LTR retrotransposons horizontal transmission is presented.

Methods and Algorithms: Fungal genomes were obtained from public databases. UniPro GenomeBrowser software (http://genome.unipro.ru/) was used for non-LTR retroelements identification and analysis. Phylogenetic analyses were performed using the MEGA 3.0 program [3].

Results: In present study we screened 34 fungal genomes for the presence of non-LTR retrotransposons. More then seventy new retroelements were found belonged to the four diverse phylogenetic clusters of elements. Phylogenetic analysis and comparisons of amino acid distances versus host divergence time which were made within and between basiodiomycete and ascomycete lineages showed the extremely low divergence between two newly identified elements from *Aspergillus niger* and *Chaetomium globosum:* It can not be explained by the strong selective constraint in the mobile elements sequence coupled with strict vertical transmission. Thus, the lower than expected divergence between elements and low rate of evolution indicate the HT event which took place less than 10 MYA.

Conclusion and Availability: The HT of non-LTR retrotransposons is extremely rare event. The actual mechanisms of HT are still unknown since it is not possible experimentally to show how the HT can occur. Parasites, symbionts, bacteria, or viruses all could be suggested as potential vectors for HT. The accumulation of new genomic data for diverse eukaryotic organisms could provide very important data for clearing up the mystery of horizontal transmission.

References:

1. V. Zupunski, F. Gubensek, D. Kordis (2001) Evolutionary dynamics and evolutionary history in the RTE clade of non-LTR retrotransposons, *Mol Biol Evol*, **18**: 1849-1863.
2. 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.
3. S. Kumar, K. Tamura, M. Nei (2004) MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment, *Brief Bioinform*, **5**: 150-163.

INFLUENCE OF AMINO ACID REPLACEMENTS ASSOCIATED WITH MULTIDRUG RESISTANCE ON β-TUBULIN MOLECULAR DYNAMICS

Nyporko A.Yu.*[*]*, Blume Ya.B.
Institute of Cell Biology and Genetic Engineering NAS of Ukraine, Kiev, Ukraine
e-mail: dfnalex@gmail,com

Motivation and Aim: Phenomenon of resistance to drugs depolymerising or irreversibly stabilising microtubules is caused by single amino acid replacements in α- and β-tubulin subunits. Most of these replacements localize immediately near binding sites of appropriate antimicrotubular compounds and, thus, can directly change their spatial structure [1]. However, spatial localization of amino acid alterations related with multidrug resistance [2, 3] doesn't correlate with localization of antimicrotubular drugs' binding sites. Thus, one could suppose that these mutations have alternative structural mechanism of action connecting with changes of behaviour of whole protein. The verification of this assumption requires of molecular dynamics analysis of tubulins, which have appropriate amino acid substitutions.

Methods and Algorithms: In accordance with literature data, different amino acid replacements in position 220 of chinese hamster β-tubulin can cause resistance to microtubule depolymerising (P->S, P->C, P->T substitutions) as well as to microtubule stabilising (P->L, P->V) drugs [3]. Spatial structures of chinese hamster β-tubulins containing these replacements and regular chinese hamster β-tubulin were reconstructed as described in our previous work [4]. Spatial structure optimization by L-BFGS method [5] and 30-ns molecular dynamics of studied proteins were calculated using the *mdrun* module of GROMACS software [6]. Structural changes were estimated by conformational energy dynamics (using *g_energy* module), levels of molecular oscillations (using *g_rms* module) and differences between conformation spaces of proteins (using *g_covar* module).

Results and Conclusion: All described replacements result in alterations of β-tubulin conformation space. Shift between average structures of regular and mutant subunits is within a range from 3.44 Å (T220 β-tubulin) to 5.51 Å (L220 β-tubulin). Four replacements P->S, P->C, P->L, P->V have cause a stable decrease of molecular oscillations level both for whole protein and amino acids forming contact surfaces between tubulin subunits. Molecular oscillations downshift is within a range from 0.72 Å (L220 β-tubulin) to 0.93 Å (S220 β-tubulin) for whole β-tubulins and from 0.69 Å to 1.06 Å for β-tubulin contact surfaces. Energy perturbations due to these four substitutions are more complicated. Replacements P220C, P220S and P220V cause essential decrease of protein conformational energy (by 570, 1144 and 572 kJ/mol correspondently), but replacement P220L decreases energy of protein by 35 kJ/mol only. However, P220L replacement causes maximal reducing of total conformational energy of contact surfaces (548 kJ/mol), though others replacements also decrease this parameter (P220C – by 336, P220S – by 267, P220V – by 392 kJ/mol). The replacement P220T influences neither level of oscillation nor conformational energy of studied proteins. Thus, we can conclude that at least four from five substitutions cause general stabilization of tubulin structure in time, that result in forming of more stable microtubules. These microtubules have non-specific resistance to any depolymerising drugs, but not to stabilising agents.

References

1. A. Yu. Nyporko, Ya. B. Blume (2006) The regularities of structural distribution in tubulin mutations responsible for responsible for resistance to antimicrotubular compounds, *Abstr. of International Symposium "The Plant Cytoskeleton: Genomic and Bioinformatic Tools for Biotechnology and Agriculture" Yalta, Ukraine, 19-23 September 2006,* P. 56-62.
2. M. Hari, Y. Wang, S. Veeraraghavan, F. Cabral (2003). Mutations in α- and β-tubulin that stabilize microtubules and confer resistance to colcemid and vinblastine, *Mol. Cancer Ther.*, **7:** 597–605.
3. S. Yin, F. Cabral, S. Veeraraghavan (2007) Amino acid substitutions at proline 220 of β-tubulin confer resistance to paclitaxel and colcemid, *Mol. Cancer Ther.,* **6**: 2798-2806
4. A. Yu. Nyporko, Ya. B. Blume (2001) Comparative analysis of secondary structure of tubulins and FtsZ proteins, *Biopolym. and Cell.* **17:** 61-69
5. B. Das, H. Meirovitch, I.M. Navon (2003) Performance of hybrid methods for large-scale unconstrained optimization as applied to models of proteins, *J Comput Chem.,* **24**: 1222-12231.
6. E. Lindahl, B. Hess and D. van der Spoel (2001) GROMACS 3.0: A package for molecular simulation and trajectory analysis, *J. Mol. Mod.* **7**: 306-317.

AGNS (ARABIDOPSIS GENENET SUPPLEMENTARY DATABASE), RELEASE 4.0

***Omelyanchuk N.A.*[1*]*, Mironova V.V.*[1]*, Zalevsky E.M.*[1, 2]*, Novoselova E.S.*[1, 2]*, Podkolodny N.L.*[1]*, Kolchanov N.A.*[1]

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

[2] Novosibirsk State University, Novosibirsk, Russia

e-mail: nadya@bionet.nsc.ru

[*] Corresponding author

Motivation and Aim: The Arabidopsis GeneNet Supplementary DataBase (AGNS) is aimed to provide an integrated view of arabidopsis development from positions of mutations in the nucleotide sequences and changes in the gene expression patterns to phenotype abnormalities [1]. AGNS contains the following sections: the Sequence database, the Reference Database, the Expression Database (ED) and the Phenotype database (PD). The current database release contains the tools for more efficient manual annotation of published papers, data processing and representation.

Methods and Algorithms: The AGNS input system was created as a Java application. The Berkeley XML DB with XQuery-based access is used for data processing.

Results: New developments at AGNS include the input system software, a redesigned web-interface with new navigation structure and automatic queries and a new application tool for the AGNS data processing (AGNK). Also the AGNS structure and format have been refined to allow accumulation a more wide spectra of data. The AGNS input system was created in order to ensure the efficient and acute annotation, storage and use of the data and, as a response to rapid growth of the content. Annotation as well as editing are facilitated in agreement with the database format and controlled vocabularies in the user-friendly interface. The following vocabularies are supported: the anatomy and morphology of arabidopsis, the developmental stages, the treatments. To facilitate the access to interrelated information in AGNS, several additional automatic queries were developed. Based on the AGNS ED, the automated queries provide the following information: (1) the key genes in the gene networks related to development of different organs; (2) the regulatory interactions between genes; (3) the changes in the gene network in the course of development. The automatic queries in AGNS PD provide information about the role of the gene or groups of genes in morphogenesis. Also, a new tool for AGNS data processing (AGNK) was created to extract information about cause-and-effect processes in gene networks related to specific developmental abnormalities [2].

Conclusion: AGNS annotates published papers on gene expression and function and by using a special format and interface integrates, systematizes, and classifies this heterogeneous, disparate, and scattered information. AGNS queries and logical data analysis in AGNK with the help of specially developed software allow retrieving complex and systematized data sets for the further investigation of different processes in morphogenesis from gene expression to functions of gene.

Availability: http://wwwmgs2.bionet.nsc.ru/agns

References:

1. N. Omelyanchuk et al. (2006) AGNS - a database on expression of arabidopsis genes, In: *Bioinformatics of Genome Regulation and Structure II.* Eds. N. Kolchanov, R. Hofestaedt, L. Milanesi, 433-442 (Springer Science+Business Media, Inc).
2. E. Zalevsky et al. (2008) AGNK: a package for knowledge mining from AGNS, *In this press*.

STATISTICAL ISSUES IN GENOME-WIDE TRANSCRIPTION FACTOR BINDING SITES ANALYSIS BASED ON CHROMATIN IP (ChIP-seq)

Orlov Y.L.*, Huss M., Vega V.B., Clarke N.D.
Genome Institute of Singapore, Singapore
e-mail: orlovy@gis.a-star.edu.sg
* Corresponding author

Motivation and Aim: Identification of transcription factor binding sites and gene regulatory elements is an important problem of computational genomics. Advances in high-throughput sequencing technologies combined with chromatin immunoprecipitation, such as ChIP-on-chip, ChIP-seq and ChIP-PET (Paired-End diTag), and the availability of human and mouse genome sequences now allow us to identify transcription factor binding sites (TFBS) and analyze mechanisms of gene regulation on the level of the entire genome [1,2]. Examples include Oct4, Sox2 and Nanog transcription factors in mouse [1] and p53, c-Myc in human. High-throughput data sets of binding sites demands integrated approaches including statistical analysis of data mapping quality, sensitivity and specificity issues and gene expression analysis [2].

Methods and Algorithms: The important bioinformatics step of ChIP-seq unlike ChIP-on-chip is the accurate mapping of short sequence reads to the reference genome [2]. The process of mapping tags to the reference genome can bias the analysis toward genomic regions with unique and complex sequence patterns, requiring adjustment of the expected chance to observe moderate peaks in ChIP sequence density. Oriented 25-mer DNA fragments obtained after Solexa sequencing were mapped and extended to 200 nt regions to count clusters of overlapping sequences (binding peaks). We develop a statistical approach based only on observed and control data to filter out noise (false positive) chromatin IP sequences. We found bias in GC content of 25-nt tags obtained after chromatin IP and in density of tags mapped onto the mouse genome (mm8, Feb 2006) related to technological and biological issues. For statistical validation of TF binding sites we used computer simulations (random genomic location of the same number of virtual sequences) taking into account background noise distribution from control (non-specific IP) data.

Results and Conclusion: Validated maps of TFBS contain thousands of experimentally defined sites and thousands f novel target genes, much more than stored in any available databases. We found strong correlations between binding motifs and cluster size. Correlation of binding affinity and sequence cluster size was shown independently for ChIP-PET experiments [1]. Combinations of different TFs maps on the same genome assembly allowed us to describe potential enhancers. We analyzed the chromosomal profile of tags and found correlation with chromatin structure and histone methylation patterns.

References:

1. Y.H. Loh, et al. (2006) The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat Genet.,* **38**(4):431-440.
2. C. Bock, T. Lengauer (2008) Computational epigenetics. *Bioinformatics*, **24**(1):1-10.

SITECON: A QUALITY TOOL FOR PREDICTION OF NEW POTENTIAL SREBP BINDING SITES. EXPERIMENTAL VERIFICATION AND ANALYSIS OF REGULATORY REGIONS OF VERTEBRATE GENES

Oshchepkov D.Y.[1]*, Ignatieva E.V.[1, 2], Vasiliev G.V.[1], Klimova N.V.[1], Merkulova T.I.[1, 2]
[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
[2] Novosibirsk State University, Novosibirsk, Russia
e-mail: diman@bionet.nsc.ru
* Corresponding author

Motivation and Aim: Methods that accurately predict transcription factor binding sites have always been important tools in studying the regulatory regions of eukaryotic genes. Here we present our algorithm SITECON implementation for SREBP binding sites recognition. Transcription factors of the SREBP family play an important role in regulation of expression of genes controlling intracellular cholesterol level and biosynthesis of fatty acids. SREBPs are functioning within the cholesterol sensor, providing gene regulation depending on the cell cholesterol level [1]. Today the list of genes regulated by SREBPs includes genes encoding enzymes of cholesterol and fatty acid biosynthesis, transport proteins, transcription factors, etc., but the list is likely incomplete. The factors of this family bind to the sites like E-box and SRE (Sterol Regulatory Element). Previously it was shown [2] that genes containing classic SREs in their promoter regions are more efficiently activated by SREBPs than those containing E-box. New SREs, found in the genes would shed a light on potential mechanisms an intact organism uses to respond to increased levels of dietary cholesterol.

Methods and Algorithms: We addressed the problem using computational approach SITECON, designed for recognition of potential transcription factor binding sites (TFBS) basing on conservative conformational and physicochemical properties detected for set of experimentally proven TFBS (3). We used EMSA with purified recombinant SREBP1-a to verify the novel SREs predicted by SITECON (4).

Results: 44 new SREs were detected in 25 out of 44 tested genes of lipid metabolism. A high predictive capacity of SITECON was proved by experimental verification: at least 20 out of tested 22 predicted sites were shown to be able to bind to SREBP1-a *in vitro* (4). As the detection threshold, SITECON employs conformational similarity (3). Accuracy of chosen threshold was demonstrated; only 3 out of 8 tested potential SREs with lower conformational similarity were shown to be able to bind to SREBP1-a *in vitro.* The discovered specific conservative properties for a set of SREs comply with the X-ray structure analysis data [5].

Conclusion: Precise method for SRE detection was developed and its accuracy was assessed using experimental data. The library of 25 genes that contain high-confidence predicted and partially tested experimentally SREs should be a valuable resource for biologists, the new genes that we have revealed as potential targets to SREBPs seem to be worthy of experimental verification for functionality.

Availability: SITECON tool http://wwwmgs.bionet.nsc.ru/mgs/programs/sitecon/.

References:

1. G.Gimpl et al. (2002) *Trends Biochem Sci.*, **27(12):**596-9.
2. M.Amemiya-Kudo et al. (2002) *J Lipid Res.,* **43(8):**1220-35.
3. D.Y. Oshchepkov et al. (2004) *Nucleic Acids Res*., **32:**208-12.
4. N.A. Kolchanov et al.. (2007) *Briefings in Bioinformatics*; **8(4):**266-274.
5. A.Párraga, et al. (1998) *Structure,* **6(5):**661-72.

FEL RADIATION USE FOR LARGE BIOMACROMOLECULES ABLATION

Peltek S.E.*[1], Goryachkovskaya T.N.[1], Dujak T.G. [1], Mordvinov V.A. [1], Kolchanov N.A., Popik V.M.[2], Scheglov M.A. [2], Kozlov A.S.[3], Malyshkin S.B.[3], Petrov A.K.[3]

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

[2]Budker Institute of Nucleic Physics SB RAS, Novosibirsk,

[3]Institute of Chemical Kinetics and Combustion SB RAS, Novosibirsk, Russia

e-mail: peltek@bionet.nsc.ru

[*]Corresponding author

Introduction and Background: We use the Budker INP free electron laser radiation with fine tuning of the wavelength from 110 μm to 240 μm for soft non-destructive ablation of various nucleic acids (phage DNA and plasmids), proteins, and enzymes.

The terahertz irradiation excites non-covalent molecular bonds (including hydrogen bonds). The radiation about 150 μm is capable to excite the out-of-the-plane deformation vibrations of the intermolecular hydrogen bonds O-H…O и O-H…N. Thus, selective dissociation of these bonds by means of FEL radiation makes it possible to transfer a biomacromolecule into the gas phase (aerosol) with the retention of the intramolecular covalent bonds. This allowed develop new method of mild and non-destructive ablation. ABLATION is defined as the removal of material from the surface of an object by vaporization, chipping, or other erosive processes. In our case ablation is transfer of biomacromolecules from solid surface into aerosol phase under FEL THz irradiation.

Results: The teraherz emission of the free electron laser was applied to the development of the technology for the biochip production standardization.

The principle of soft nondestructive ablation of biological macromolecules under terahertz irradiation was applied to technology for the direct analysis of the target DNA from biochip surface. There was prepared model biochip for this experiment. DNA-probe is covalently bonded to biochip surface. After hybridization the target DNA was bonded to the DNA-probe by hydrogen bonds. By the action of terahertz emission hydrogen bonds were disrupted and target DNA was transferred into aerosol phase. The target DNA was collected to the filter for subsequent analysis. Firstly, ablated target DNA was amplified by Polymerase Chain Reaction. Polymerase Chain Reaction product was identified by electrophoretic analysis. The electrophoretic analysis proved identity of initial and ablated DNAs. Secondary, the following sequence analysis proved the sequence identity of initial and ablated DNAs.

Conclusion: By action of terahertz emission we can destroy hydrogen bonds, leave covalent bonds intact and transfer large biomacromolecules into aerosol phase.

EXPRESSION PROFILING USING SECOND GENERATION SEQUENCING TECHNOLOGIES

Parkhomchuk D.**, Banaru M., Borodina T., Amstislavskiy V., Soldatov A., Lehrach H.
Max-Planck Institute of Molecular Biology, Berlin, Germany
e-mail: parkhomc@molgen.mpg.de
* Corresponding author

Motivation and Aim: Evaluation of second generation sequencing platforms for whole-genome expression profiling.

Methods and Algorithms: Second generation sequencing platforms offer new opportunities for transcriptome analysis. Sequencing approach has significant advantages over conventional hybridization-based (microchip) methods: (i) no preliminary information about transcriptome is required; (ii) any desirable detalization of the analysis is possible; (iii) new transcripts may be detected; (iv) comparison of results obtained in different laboratories becomes easy and straightforward. Transcriptome sequencing is also more informative than SAGE-analysis, because it gives information about internal structure of transcripts.

Results: Here we report technological improvements for expression profiling studies and for transcript structure analysis. Combination of random sequencing approach with paired-end technology allows us to get information about structure and overall size of transcripts. With novel protocols we are able to derive information about transcript direction, which is useful for resolution of complementary transcripts and exons prediction. We demonstrated the ability to provide genome-wide allele-specific expression profiling which reveals parental-specific (imprinting) and allele-specific biases. We also present a bioinformatics pipeline for transcriptome analysis using Illumina sequencing platform data.

Conclusion: Second generation sequencing platforms demonstrate numerous advantages over other expression profiling methods and are likely to outfit older methods in future. The examples we provided show some unique features of the platform and new methods of data analyses.

Availability: additional information is available upon request

DETAILED STATISTICAL ANALYSIS OF AROMATIC INTERACTION IN PROTEINS

Pereyaslavets L.B.*
Moscow Institute of Physics and Technology, Moscow, Russia
e-mail: pereyaslavets.l@gmail.com
* Corresponding author

Motivation and Aim: Investigation of aromatic structures interaction in proteins is one of ways to understand how they form their tertiary structure. [1] More precise than previous analysis of aromatic stacking [1,2] can reveal new insight of necessity and functioning of aromatic amino acids in proteins. Comparing of aromatic dimers frequencies to ab initio high quality dimer energies [3] under the assumption of quasi-Boltzmann distribution [4] can shows similarity and difference of aromatic amino-acids dimers arrangements in proteins, compared to those in vacuum.

Methods and Algorithms: For all statistical investigations of studied dimer conformations unredundant set of protein chains from PDB (about seven thousand) was used. From this set all aromatic pairs was picked out and bring to one base (first dimer element) Quantity of second mobile dimer element mass center in unit of volume was used as population in quasi-Boltzmann law. Because of obtained data limit there is only consideration of a various layers of data in all 6D solid residues space.

Results: After comparing of logarithm of population to quantum energies of high quality by quasi-Boltzmann distribution for the first time was obtained "experimental" energies (based on population) of parallel stacking of PHE-PHE, PHE-TYR, TYR-TYR that contrast to quantum ones. There is full correlation for PHE-PHE parallel stacking with effective distribution temperature between 300-400K, partial coincidence for other parallel dimers. It turned out, that it is impossible to reproduce quantum energy difference between Parallel Displaced and T-Stacking (for stacking definitions see ref. [3]) because of impossibility of obvious taking into account addition interaction of surrounding groups.

Detailed dependencies from various physical variables (distances between aromatic planes, angle between planes ant other) for PD and T stacking were obtained and it corresponds to quantum picture with according to physical restriction of aromatic residues.

Conclusion: Aromatic stacking in general obeys the quasi-Bolztmann law. The size of aromatic groups in line (PHE, TYR, TRP) make a great impact on this distribution by influence of rigid group scale, presumably on protein folding stage. It's another one prohibitive force on way of proteins evolution and diversity that must be accounted for protein structure prediction without explicit folding modeling.

References:

1. S.K.Burley, G.A.Petsko (1985) Aromatic-Aromatic Interaction: A Mechanism of Protein Structure Stabilization, *Science*, 229: 23-28.
2. G.B.McGaughey, M.Gagne, A.K.Rappe (1998) π-Stacking interactions. Alive and well in proteins, *J. Biol. Chemistry,* 273: 15458-15463.
3. S.Tsuzuki, K.Honda, T.Uchimaru, M.Mikami (2005) Ab initio calculations of structures and interactions energies of toluene dimers including CCSD(T) level electron correlation correction, *J. Chem. Phys,* 122: 144323.
4. A.V.Finkelstein, A.Y.Badretdinov, A.M.Gutin (1995) Why do protein architectures have Boltzmann-like statistics? *Proteins: Structure, Function, and Genetics,* 23(2): 142-150.

CPG ISLANDS EVOLUTION: CPG DINUCLEOTIDES DEATH AND BIRTH PROBABILITIES IN DIFFERENT GENOME REGIONS

Pertsovskaya I.[1*], Oparina N.[2], Vinogradov D.[3], Favorov A.[2, 4], Mironov A.[1, 3]
[1] Moscow State University, Moscow, Russia
[2] GosNIIGenetica, Moscow, Russia
[3] IITP RAS, Moscow, Russia
[4] Sidney Kimmel Cancer Center at Johns Hopkins, Baltimore, MD, USA
e-mail: inna.perts@gmail.com
[*] Corresponding author

Motivation and Aim: Cytosine methylation is the only known covalent modification of DNA. It plays critical role in gene expression regulation, cell development, imprinting etc. The cytosine (C) methylation occurs in CpG dinucleotides, so the rate of CpG pairs is lower than the independent probabilistic prediction. Yet, there are specific genome regions that are enriched with CpG dinucleotides. These areas are referred to as CpG islands. They are usually unmethylated. One of directions of epigenomic studies is the investigation of the CpG islands evolution. Thus, the work we present here examines the mutations of cytosine in CpG context. Also, we describe orthologous CpG islands, i.e. CpG islands located into or near (in upstream or downstream) orthologous genes and also look for highly similar islands.

Methods and Algorithms: We used triple alignment of human, chimpanzee and macaque genomes and of human, rat and mouse genomes to analyze the mutations of cytosine in CpG dinucleotides. Multiple alignments are available at http://hgdownload.cse.ucsc.edu/goldenPath/hg18/multiz28way We used annotaded CpG islands from Genome Browser [1] and found by CpGcluster algorithm [2]. Also we did the whole genome analysis of CpG dinucleotides births and deaths. The sequence similarity of highly similar island was at least 90% according to BLAT [3]. The orthologous genes positions were obtained from Genome Browser [1]. All the scripts were written in Perl.

Results: We showed that the mutations notCpG→CpG are rarer than CpG→notCpG both in and outside CpG islands. The mutations CpG→notCpG inside CpG islands are rarer than outside, while the backwards changes notCpG→CpG are more often inside islands. Also we have identified 220 groups of orthologous CpG islands and 434 islands combined in 158 groups of high similarity.

Conclusion: The objects that are commonly referred as CpG islands have complex nature and possibly there is a set of different biological and statistical phenomena combined into one term. On the other hand, at least some of the objects have biological rather than only statistical nature.

Availability: Perl scripts are available on request from the authors.

References:

1. Kent, W.J., Sugnet, C. W., Furey, T. S., Roskin, K.M., Pringle, T. H., Zahler, A. M., and Haussler, D. (2002). The Human Genome Browser at UCSC. Genome Res. **12**(6), 996-1006.
2. Hackenberg, M., C. Previti, et al. (2006). CpGcluster: a distance-based algorithm for CpG-island detection, *BMC Bioinformatics,* **7**: 446.
3. Kent, W.J. (2002) BLAT - The BLAST-Like Alignment Tool. Genome Res. **12**(4), 656-664.

ASSOCIATION OF RNA STRUCTURES AND SPLICING

Pervouchine D.D.[1], Raker V.A.[2], Gelfand M.S.[1, 3], Mironov A.A.[1*]
[1] Department of Bioengineering and Bioinformatics, Moscow State University, Moscow, Russia, e-mail: mironov@boiinf.fbb.msu.ru
[2] Centre de Regulacio Genomica, Barcelona, Spain
[3] Institute for Information Transmission problems RAS, Russia

Motivation and Aim: RNA structure has been reported to impact many cellular processes, including splicing in genes associated with disease . The question is: can RNA structure have an impact on splicing.

Methods and Algorithms: We searched possible RNA structures that can loop-out introns. We define putative event of intron looping-out if the region near donor contains a word that have a complement near acceptor site (word size at least 9 nt). For all found events we analyze if the complement word are conservative in relative genomes. Three sets of genomes was analyzed: vertebrates, *Drosophilae*, worms. For some cases the experimental validation was done.

Results: In every set of genomes about 200 genes were selected. The distribution of position of complement blocks is not uniform; alternatively spliced introns are overrepresented; long introns are overrepresented. Gene ontology analysis shows that selected genes mostly are related (with high significance) to follow categories: cell development, DNA binding, neurons development, signal cascades etc. Experimental validation was based on minigene system and was provided in follow way. A mutations in one found box were introduced. Then a compensatory mutations in complement box were introduced. The splice forms were tested using RTPCR and sequencing. In all cases mutations in one box significantly changes pattern on splicing and compensatory mutations rescue the wild type splicing.

Conclusion: RNA structures play an important role in splicing. Seems our observations support hypothesis that RNA structure may be involved in the regulation of splicing.

TRANSCRIPTOME ANALYSIS OF ERF1-MEDIATED NITROGEN SIGNALING

Petrova A.V.[1], Dagkessamanskaya A.[2], Trouilh L.[3], Labourdette D.[3], Sokol S.[3], Francois J.M.[2, 3], Zhouravleva G.A.[1*]

[1] St.Petersburg State University, Department of Genetics, Saint-Petersburg, Russia
[2] Centre de Bioingénierie Gilbert Durand UMR-CNRS 5504, UMR-INRA 792, Institut National des Sciences Appliquées F-31077, Toulouse Cedex 04, France
[3] Transcriptome-Biochips Platform of Genopole Toulouse Midi-Pyrénées, France
e-mail: zhouravleva@rambler.ru
* Corresponding author

Motivation and Aim: Despite the profound studies in the field of translation termination, many aspects of this phenomenon remain unclear. The good example of such situation is a pleitropic manifestation of mutations in essential gene *SUP45* encoding translation termination factor 1 (eRF1) in the yeast *Saccharomyces cerevisiae*. Mutations in *SUP45* gene cause various defects in the cell. It was shown in our laboratory that *sup45* mutants lose capability to undergo the pseudohyphal transformation. In this work, we performed a transcriptome analysis for the purpose of examining the changes of expression patterns in the presence of wild-type or mutant allele of *SUP45* gene under the condition of nitrogen limitation.

Methods and Algorithms: Two diploids strains with different genetic backgrounds were used in this work. Strain D1667 originates from Peterhoff Genetic Collection. Strain D1643 comes from Berkley collection and is characterized by Σ1278b background. Strains, carrying wild-type or mutant (*sup45-103*) allele of *SUP45* gene, were grown simultaneously in liquid YNB and SLAD media. Overnight cultures were diluted and grown to the OD_{600}~0,6 and then used for total RNA extraction using Qiagen RNeasy Mini Kit. Purified RNA was subjected to reverse transcription, labeled with cyanine 3-dCTP and cyanine 5-dCTP dyes and used for transcriptome analysis (http://biopuce.insa-toulouse.fr/protocoles.php). Two independent experiments were made for each variant. The obtained slides were analyzed using BioPlot software (http://biopuce.insa-toulouse.fr/ExperimentExplorer/doc/).

Results: Analysis revealed only few genes, which expression was changed in both mutant strains: *ENA2* (ion transport), *DAL80* (regulation of nitrogen utilization), *VPS29* (retrograde transport endosome to Golgi), *HSP82* (protein folding), *GPI1* (GPI anchor biosynthetic process) and others. Only *DAL80* gene directly participates in regulation of nitrogen metabolism. In mutant strains its expression increases 1,25 and 1,4-fold more than in D1643 and D1667 wild-type strains, respectively. It is possible that in *sup45-103* strains increased expression of *DAL80* may influence expression of other genes, sensitive to nitrogen catabolite repression, and thus change the sensing of quality and availability of nitrogen source. This may, in turn, block the expressional and morphological changes, caused by nitrogen starvation. Considering background difference, D1643 strain demonstrates more pronounced sensitivity to nitrogen starvation than D1667 strain. Expression of *STE11*, coding for signal transducing MEK kinase essential for pseudohyphal and invasive growth, is lower in D1667 strain comparing to D1643 strain, and in D1643 *sup45-103* comparing to wild-type strain.

This work was supported by FEMS Research Fellowship (for AP) and RFBR grant № 07-04-00605-a (for AP and GZ).

MOLECULAR DYNAMICS SIMULATION OF ATP BINDING BY THE M. TUBERCULOSIS PROTEIN PII

Pintus S.S.[1]*, Ramachandran S.[2], Ivanisenko V.A.[1]
[1]Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
[2]Functional Genomics Unit, Institute of Genomics and Integrative Biology,
Delhi 110 007, India
e-mail: pintus@bionet.nsc.ru
* Corresponding author

Motivation and Aim: The nitrogen regulatory protein pII is involved in nitrogen metabolism in prokaryotes, particularly in *Mycobacterium tuberculosis*, which is a pathogenic organism of tuberculosis. pII indirectly activates the transcription of the *glnA* gene, that encodes the glutamate synthetase enzyme, which catalyses the reaction of nitrogen assimilation. It has been demonstrated by Ramachandran group that pII is a nucleotide binding protein[1]. Molecular dynamics simulations were performed to estimate the free energy of ATP binding to pII, conformationally modified by 2-ketoglutarate binding.

Methods and Algorithms: The tertiary structure of pII complexed with ATP and modified by 2-ketoglutarate binding was taken from the previous paper [1]. Then, the GROMACS molecular topology and atom charges were obtained for the ligand using the PRODRG2 web server. Then, the molecular dynamics simulations were performed, including energy minimization, protein relaxation, molecular dynamics itself and refinement.

Results: The ATP kept its position in its binding pocket and the conformation of the loop that holds 2-ketoglutarate remained unchanged (bended) during all simulation period. Nevertheless, opening and closing of hydrogen bonds was observed. The T-loop residue R38 also formed a hydrogen bond with ATP.

Conclusion: The evidence was obtained, that the pII protein, which conformation is modified by the 2-ketoglutarate binding, strongly binds ATP. The T-loop residue R38 is supposed to be involved in ATP binding, and observation of the hydrogen bond during the simulations provides extra evidence for that.

References:

1. A.Bandyopadhyay et al. (2008) Molecular characterization of the PII protein of *Mycobacterium tuberculosis* reveals that it is a general nucleotide binding protein, *Journal of Biochemistry,* in press

COEVOLUTION OF PROTEIN DOMAINS OF P53 AND MDM2 – KEY PROTEINS OF APOPTOSIS

Pintus S.S.*, Ivanisenko V.A.
Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
e-mail: pintus@bionet.nsc.ru
* Corresponding author

Motivation and Aim: The tumor suppressor p53 protein induces cell cycle arrest and apoptosis through its transcription factor activity as well as through protein-protein interactions (Levine, 1997). Interacting protein regions are considered to undergo coevolution that is usually revealed in protein distance correlation analysis (Pazos et Valencia, 2001). Here we study coevolution of the p53 domains with domains of proteins that are involved in the p53 gene network. Also we analyze the effect of exclusion of neutrally evolving codons on correlation of protein distances to study the phenomenon of high correlation values for noninteracting proteins that is considered to result from neutral evolution.

Methods and Algorithms: Protein sequences of p53 transcription activation domain (TAD), p53 DNA binding domain (DBD) and mdm2 SWIB domain were for eight vertebrate species were taken from the Swiss-Prot database. Corresponding coding sequences were taken from the Genbank database. Protein sequences were aligned using the ClustalW program and corresponding codon alignments were obtained. Protein distance matrices were obtained using the LAPD program (http://www.csc.kth.se/~arve/code/lapd/), developed by Lars Arvestad (forthcoming paper). The Pearson correlation coefficient between the matrices was calculated using the *ad hoc* Perl program. We used the PAML software to search for neutrally evolving codons.

Results: Correlations of protein distances were as follows: p53 TAD-mdm2 SWIB – 87%, p53 DBD-mdm2 SWIB – 86%, p53 TAD-p53 DBD – 50%. The proportions of neutrally selected codons were: mdm2 SWIB – 0%, p53 DBD – 14%, p53 TAD – 60%. After exclusion of neutrally selected codons the correlation values slightly decreased: p53 TAD-mdm2 SWIB – 85%, p53 DBD-mdm2 SWIB – 79.5%, p53 TAD-p53 DBD – 43%. Interestingly, the correlation between neutrally evolved regions of p53 TAD and p53 DBD was 82%.

Conclusion: The mdm2 SWIB domain, that inhibits p53, appears to be the most conserved of the three peptides under study, while its target p53 TAD domain exhibits high variability. This result suggest the strongest purifying selection for the domain that supplies a negative feedback on p53 concentration in cell. Low correlation between "regulator" p53 TAD and "effector" p53 DBD suggests their independent evolution, which may result from adaptation of the p53 TAD to new protein signaling during evolutionary history.

References:

1. A.Levine (1997) *Cell*, **88**:323-331.
2. F.Pazos, A.Valencia (2001) *Protein Eng*., **14**:609-614.

THE APPLICATION OF SITEGA AND OPTIMIZED PWM METHODS FOR CLOCK/BMAL BINDING SITES RECOGNITION

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

Motivation and Aim: Circadian rhythms are near-24-hour rhythms which is extremely important for essence of living systems. Transcription factor Clock/Bmal is, a master regulator of circadian rhythm. Here we present detecting of potential Clock/Bmal binding sites (BSs) in the regulatory regions of genes with circadian cycling expression pattern.

Methods and Algorithms: Nucleotide sequences of Clock/Bmal BSs were retrieved from TRRD [1]. The combination of SiteGA and oPWMs (optimized position weight matrix) models was used for putative sites prediction [2]. The SiteGA method applied genetic algorithm to infer specific set of locally positioned dinucleotides (LPDs). To set thresholds we fixed stringencies corresponding to recognition of certain portions of train data (50-90%).

Results: Dinucleotide matrix captured comparatively long flanking sequences, so that total length of each sequence was 38 nt. SiteGA method allowed us to find out a number of significant correlations between frequencies of LPDs. Notably, they were found not only in the canonical consensus region, even the most significant correlations distributed along the full length of analyzed sequence. Two sample of sequence were compiled for potential BS search. The first samples contained [5000;+1] upstream regions of 109 mouse genes for which circadian expression was confirmed by high-density oligonucleotide array-based analysis [3]. These genes were subdivided onto 8 functional groups according to their participation in certain specific for liver or suprachiasmatic nuclei functions. Potential BSs of Clock/Bmal were not found for each analyzed sequence, but for any groups they were found. The second sample contained first 5000 upstream and whole transcript regions of 9 human genes involved in heme biosynthesis, also this sample was supplemented with ortologous mouse, rat and cattle genes. In mammals heme biosynthesis is controlled by circadian clock through regulation of rate limiting enzyme ALAS1 (or ALAS2 in erythroid cells). We predicted potential BSs in the upstream regions of human, mouse and cattle ALAS2 genes, and in the last introns of human, mouse and rat ALAS1 genes. Potential Clock/Bmal BSs in the respective regions (upstream or intron) of 3 ortologous genes argues for detection of most promising targets.

Conclusion: Analysis of potential Clock/Bmal BSs indicated that only restricted subset of genes with circadian cycling expression may be considered as promising targets of this factor. In particular, these genes may be competent for interactions of circadian oscillator with transcription networks maintaining tissue and organ specific functions.

References:

1. N.A. Kolchanov et al. (2002) Transcription Regulatory Regions Database (TRRD): its status in 2002. *Nucl Acids Res*, **30:** 312-317
2. V.G. Levitsk**y** 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. Panda et al. (2002) Coordinated transcription of key pathways in the mouse by the circadian clock. *Cell,* **109:** 307-320.

A DATABASE FOR ANALYSIS OF THE ORGANIZATIONAL FEATURES OF THE PROMOTER REGIONS IN THE CO-EXPRESSED GROUPS OF GENES

***Podkolodnyy N.L.*[1, 2, 3]*, *Ignatieva E.V.*[1, 2], *Nechkin S.S.*[1, 2], *Ananko E.A.*[1], *Podkolodnaya O.A.*[1]**

[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia;
[2] Novosibirsk State University, Novosibirsk, 630090, Russia;
[3] Institute of Computational Mathematics and Mathematical Geophysics, SB RAS, 630090, Russia
e-mail: pnl@bionet.nsc.ru
* Corresponding author

Motivation and Aim: In multicellular eukaryotic organisms, the transcriptional activity of a particular gene is dependent of the type of cell, organ, or tissue, developmental stage of the organism, cell cycle, or cell differentiation, also on numerous inducers or repressors, etc. This fine and complex regulation is provided by the involvement of a great variety of regulatory proteins and mechanisms underlying their functioning. The aim of this work is to develop the RETRA database designed to analyze and reconstruct of the mechanisms of tissue-specific regulation of gene transcription.

Methods and Algorithms: The RETRA database has been developed on the basis of an integration of data from various world resources, and it contains the following information:

i) the structural-functional organization of the gene transcription regulatory regions (TRRD) [1]; ii) gene localization in whole genomes (EntrezGene, RefSeq); iii) estimation of gene expression level in different tissues and organs (UniGene, GEO); iv) functional gene annotation (Gene Ontology).

RETRA contains data on the localization of the regulatory regions and TFBSs in whole genome. The expression levels of genes in different cells, tissues, organs at different developmental stages of cells, tissues and the organism were estimated on the basis of analysis of EST and microarray data.

The database contains information about the morphological characteristics and the functional state of the cells that express genes. For example, it includes information about the TFs expressed in cells and responsible for tissue-specific expression.

The database includes also a description of the regulatory patterns, knowledge about the mechanisms regulating transcription, about the role of the TFs and other proteins in transcription regulation retrieved from scientific publications.

Conclusion: The RETRA database designed to resolve the following issues was developed:

i) Search of a set of co-expressed genes.
ii) Functional annotation of a set of co-expressed genes.
iii) Search of organizational patterns of the promoters of co-expressed genes.
iv) Reconstruction of the tissue-specific mechanisms of transcription regulation
v) Prediction of the relative level of gene expression in different tissues

References:

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

MECHANISMS OF COMMUNICATION OVER A DISTANCE ON DNA AND CHROMATIN

Polikanov Y.S., Studitsky V.M.
Graduate School of Biomedical Sciences
Department of Pharmacology, University of Medicine and Dentistry of New Jersey (USA)

Motivation and Aim: Regulatory elements in eukaryotic cell nuclei can specifically and efficiently interact with their targets over long stretches of chromatin-compacted DNA, but the mechanisms of communication in chromatin remain poorly investigated. Our initial studies suggest that efficient communication on chromatin templates *in vitro* may involve uncoiling of nucleosomal DNA and/or internucleosomal interactions [1].

Methods: The mechanism of communication in chromatin was addressed using physiologically relevant saturated arrays of precisely positioned nucleosomes [2].

Results: Intact histone N-terminal tails are essential for efficient communication on saturated, but not sub-saturated arrays. Chromatin supports efficient communication over distances from 0.6 to at least 5 kb. The arrays organized in 10-nm or 30-nm chromatin fibers can support communication *in cis* with similar efficiencies, suggesting that both transient nucleosomal DNA uncoiling and dynamic, close-range internucleosomal, intrafiber interactions mediated by histone N-tails (brachiation) are involved in the communication. Communication between chromatin fibers organized in different topological domains occurs much less efficiently suggesting that interdomain brachiation is prevented by the loop formation.

Conclusion: Thus chromatin is a highly dynamic and flexible structure that allows efficient intradomain communication between distant regulatory DNA elements. We also found that the rate of bridging in the system that we used is not limited by diffusion of the DNA sites one to another and is determined by the equilibrium probability of juxtaposition of the DNA sites [3]. We argue that this conclusion derived for the specific system is likely to be valid for the great majority of biological processes involving protein-mediated DNA looping.

References:

1. Rubtsov M. A.*, **Polikanov Y. S.***, Bondarenko V. A., Wang Y. H. and Studitsky V. M. (2006) Chromatin structure can strongly facilitate enhancer action over a distance. *Proc Natl Acad Sci USA* **103:** 17690-5. * - equal authorship
2. **Polikanov Y. S.**, Rubtsov M. A. and Studitsky V. M. (2007) Biochemical Analysis of Enhancer-Promoter Communication in Chromatin. *Methods* **41(3):** 250-8.
3. **Polikanov Y. S**., Bondarenko V. A., Tchernajenko V., Jiang Y. I., Lutter L., Vologodskii A. and Studitsky V. M. (2007) Probability of the Site Juxtaposition Determines the Rate of Protein-Mediated DNA Looping. *Biophys J* **93(8):** 2726-31.

QUALITY OF LOCAL AND GLOBAL PAIR-WISE ALIGNMENTS OF AMINO ACID SEQUENCES

***Polyanovsky V.*[1], *Roytberg M.**[2], *Tumanyan V.*[1]
[1] Institute of Molecular Biology, Moscow, Russia
[2] Institute of Mathematical Problems of Biology, Pushchino, Moscow Region, Russia
* Corresponding author: e-mail: mroytberg@impb.psn.ru

Motivation and Aim: In many applications, the algorithmic alignment of two protein sequences ideally should restore the genuine one, i.e. the alignment that superimposes positions originating from the same position of the common ancestor of the proteins. Thus, it is important to know the average accuracy and confidence of the algorithmic alignments depending on the sequence similarity and to understand when the global alignment leads to better results than the local one and *vice versa.*

Methods and Algorithms: The local and global versions of Smith-Waterman alignment algorithm with standard values of parameters were considered. We have performed computer experiments; each experimental set consists of 1000 pairs of artificial amino acid sequences. Given the set, we align locally and globally all its pairs, and then compute average values of alignment accuracy and confidence. Each sequence pair S_1, S_2 in the experiment was generated with the following procedure, depending on three parameters ***(P, a, b);*** the parameters are the same for all sequences of the set**.** First, a random amino acid sequence C_0 of length $\boldsymbol{L}$ = 200 was generated. Second, we have produced two its independent "ancestors" C_1 and C_2 using the evolutionary model [1]; the model allows both replacements and indels, the frequencies of mutation events depend on the PAM value ***P.*** Third, we independently generate random "wings" B_1, B_2, E_1, E_2, the sequences to be compared are $S_1 = B_1 \cdot C_1 \cdot E_1$ and $S_2 = B_2 \cdot C_2 \cdot E_2$. The genuine alignment of S_1 and S_2 superimposes only positions of the "cores" C_1, C_2 corresponding to the same position of C_0. The wings reflect that S_1 and S_2 may have only local similarity. The lengths of the wings meet two conditions: (1) the total length $|B_1|+|E_1| = |B_2|+|E_2| = \boldsymbol{a \cdot L}$; (2) the cores Ci within Si (i=1, 2) are shifted into different directions: $|B_1| = |E_2| = (1- \boldsymbol{b}) \cdot (|B_1|+|E_1|)/2$ ($\boldsymbol{b}$ = 0 corresponds to the absence of the shift). We have checked all combinations (***P, a, b)*** where ***P*** = 30, 60, 120, 240 (that is ~ 60, 40, 20 and 10 %id)***; a***= 0.1, 0.2, 0.5, 1.0, 2.0; ***b***= 0, 0.1, **…,** 0.9, 1**.**

Experiments and results: (1) Average confidence and accuracy are almost equal both for local and global alignments in all experiments. (2) Given ***P*** and ***a,*** the accuracy of the global alignment is almost equal to that with ***b*** =0 or is almost zero (as for ***b*** =1). The cut-off value of ***b*** depends on ***P*** and ***a*** **(**e.g. for ***P*** = 120 the cut-off is 0.1 if ***a*** =2.0; 0.2 for if ***a*** =1.0; 0.4 if ***a*** =0.5). (3) If ***b*** =0, then the global alignment has better accuracy than the local one. This suggests the 2-step alignment procedure: (1) perform *local* alignment to find similar fragments F_1, F_2 of the compared sequences S_1, S_2; let $F_i = S_i[\boldsymbol{c_i}, \boldsymbol{g_i}]$; $\boldsymbol{d_i = g_i - c_i}$. (2) perform *global* alignment of enlarged fragments $H_i = S_i[\boldsymbol{c_i - k d_i}, \boldsymbol{g_i + k d_i}]$ where ***k*** depends on sequence similarity, e.g. ***k***=0.5 if ***P=120***. The computer experiments show that the procedure improves the accuracy of local alignment on 5-10% , e.g. for ***P=120; a = 1; b=0.6.*** it gives 69 % instead of 64%

Conclusion: The global alignment gives better accuracy than the local one if the similar fragments are placed in the same positions of their sequences even if the non-similar wings are as long as the similar cores. This observation leads to the improvement of the accuracy of local alignment.

Reference:

1. Benner,S.A., Cohen,M.A. and Gonnet,G.H. (1993) Empirical and structural models for insertions and deletions in the divergent evolution of proteins. *J. Mol.Biol*., 229, 1065-1082.

MOSAIC NATURE OF THE WATER-LIPID INTERFACE AFFECTS A BEHAVIOR OF MEMBRANE-ACTIVE PEPTIDES

Polyansky A.A.*, Volynsky P.E., Efremov R.G.
M.M. Shemyakin and Yu.A. Ovchinnikov Institute of Bioorganic Chemistry Russian Academy of Sciences, Moscow, 117997, Russia
e-mail: newant@gmail.com
* Corresponding author

Motivation and Aim: According to modern conceptions, the cell membrane possesses mosaic and multilevel organization with the nanosized heterogeneous in-plane distributions of lipids, along with their structural and dynamic properties. Here we scrutinize a phenomenon of a mosaic polarity nature of the membrane-water interfaces. Previously, it was shown that the structural plasticity of some membrane active peptides (MAPs) determine their complicated behavior upon binding to model membranes [1]. A reasonable explanation of this effect can be achieved if the mosaic hydrophobic/hydrophilic nature of the membrane surface is taken into consideration.

Methods and Algorithms: The role of microscopic properties of the water-lipid interface is investigated *via* computer simulations of penetratin (cell-penetrating MAP) in zwitterionic (DOPC) and anionic (DOPS) model membranes. To describe polar nature of the water-lipid interface, we proposed to calculate and map the molecular hydrophobicity potential (MHP) on the surfaces of explicit hydrated lipid bilayers. Previously, the MHP-approach has been successfully used to characterize polarity properties of membrane proteins and MAPs [2].

Results: From the hydrophobicity point of view, the solvent-accessible surfaces of hydrated lipid bilayers possess a prominent "mosaic" character. This is reflected in occurrence of dynamic clusters of hydrophobic surface area created by acyl chains of lipid molecules exposed on the interface. Such "mosaic patterns" are specific for lipid bilayers of a particular composition. In the DOPS membrane, they determine initial stages of penetratin adsorption, which strongly depend on the "complementarity" between polarity properties of the peptide and its local interfacial environment. In the case of high complementarity, the peptide penetrates deeply into the membrane without significant destabilization of its initial secondary structure. An alternative scenario demonstrates partial unfolding of the peptide on the interface in order to compensate unfavorable peptide-membrane interactions. Thus, depending on the overall and microscopic properties of the water-lipid interface, penetratin is capable of utilizing different pathways to realize its biological activity against cell.

Conclusion: To summarize, we should outline the following: i) Heterogeneous distribution of hydrophobic and hydrophilic properties on the surface seems to be an inherent feature of a particular lipid bilayer, which depends on the chemical structure of lipid polar heads. ii) Upon binding of MAPs to the water-lipid interface, the correspondence between polarity patterns of peptide and membrane surfaces plays an important role and may explain (at least partially) the complicated behavior of MAPs.

References:

1. Polyansky A.A., Volynsky P.E., Efremov R.G. (2007) Computer simulations of membrane-lytic peptides: perspectives in drug design. *J. Bioinform. Comput. Biol.,* **5**: 611-626.
2. Efremov R.G., Chugunov A.O., Pyrkov T.V., Priestle J.P., Arseniev A.S., Jacoby E. (2007) Molecular lipophilicity in protein modeling and drug design. *Curr. Med. Chem.,* **14**: 393-415.

ARABIDOPSIS THALIANA miRNA ABUNDANCE RANGE CORRELATES WITH THE TBP/TATA-AFFINITY OF microRNA GENES

Ponomarenko P.M.[1,*], Ponomarenko M.P.[2], Omelyanchuk N.A.[2], Kolchanov N.A.[1,2]
[1] Novosibirsk State University, Novosibirsk, Russia
[2] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia.
e-mail: pon@bionet.nsc.ru
[*] Corresponding author

Motivation and Aim: The miRNA abundance in any cell depends on (i) the level of transcription and (ii) the degradation rate. Previously we identified [1] that at the mature miRNA sequences, the specific WRHW and DRYD tetramers are significantly associating with miRNA abundance in plant tissues. With this in mind, the intrigue question raises what is a vital motivation and aim for miRNA gene transcription regulation?

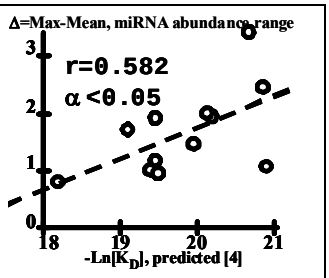
Methods and Algorithms: To answer this question we compared the miRNA abundance data [2] with the data on miRNA gene transcription [3]. We selected 12 families detected in every microarray probe among miRNA data [2]. Then we represented each family by the most frequently observed gene [3]. Finally, we calculated -Ln(K_D), TBP-affinity rate to selected gene promoter by using its known TATA-box location [3], as given elsewhere [4]:

$$-Ln(K_D) = 10.9 - 0.23 \times Ln(K_{D,2}) + 0.15 \times PWM - 0.2 \times Ln(K_{D,1}).$$

Results and Conclusion: TBP/TATA-affinity appears significantly correlating with miRNA abundance range, Δ=Max-Mean (Table). This allows us to conclude that the transcription regulation defines miRNA abundance range. Within this range, the cell-specific miRNA abundance rate can be adjusted by miRNA transportation and functioning machinery [1].

Table. The comparison between miRNA abundance [2] and transcription [3] by TBP/TATA-affinity, Ln[K_D] [4]

miRNA abundance [2]			miRNA transcription [3]			-Ln[K_D] [4]
Family	Mean	Max	Gene	Found/Tested	TATA-box	
miR159	5.24	6.93	b	5/6	TTTAAAAA	19.11
miR160	3.87	5.80	c	6/6	TATATATT	20.20
miR163	1.75	5.14		5/7	TATAAATA	20.68
miR164	4.19	5.26	a(2)	5/6	TATATATA	20.91
miR165	0.88	1.88	a	5/6	TATTAAAA	19.40
miR166	1.49	2.29	b	5/6	TTAAAAAC	18.19
miR167	5.06	6.51	b	3/6	TATCTATA	19.96
miR170	1.33	2.27		4/5	TTATTAAA	19.50
miR172	5.92	7.89	e(1)	8/9	TATAAAAG	20.18
miR394	2.08	3.26	a(1)	4/5	TATTAAAA	19.46
miR396	4.17	6.09	a	5/6	TATAAATA	19.47
miR398	1.22	3.67	c	3/6	TATATATA	20.87



TBP-affinity [4] to TATA-box [3] correlates significantly to miRNA abundance range [2].

References:

1. M. Ponomarenko et al. (2008) The content of microRNAS in *Arabidopsis thaliana* correlates with the occurrence of WRHW and DRYD tetramers. *DAN* (In press).
2. M.J. Axtell and D.P. Bartel (2005) Antiquity of microRNAs and their targets in land plants. *Plant Cell*. **17**:1658-1673.
3. Z. Xie et al. (2005) Expression of Arabidopsis MIRNA genes. *Plant Physiol.* **138**:2145-2154.
4. P. Ponomarenko et al. (2008) A stepwise model of TBP/TATA box binding allows for predicting human hereditary diseases by single nucleotide polymorphism. *DAN*, **419**:88-92.

THE PRECISE EQUILIBRIUM EQUATION OF TBP/TATA-BINDING PREDICTS HUMAN FAMILIAL DISEASES UPON MUTATIONS

Ponomarenko P.M.[1], Savinkova L.K.[2], Drachkova I.A.[2], Lysova M.V.[2], Arshinova T.V.[2], Ponomarenko M.P.[2]*, Kolchanov N.A.[1, 2]

[1] Novosibirsk State University, Novosibirsk, Russia.

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

e-mail: pon@bionet.nsc.ru

[*] Corresponding author

Motivation and Aim: During the Post Genome Era, the computer-based calculation on how mutations can alter human health is one of the most challenging problems in bioinformatics.

Methods and Algorithms: Based on the newest magnitudes [1] of TBP-affinity to native TATA-boxes, our cross-validation tests of Bucher's matrix [2] and relationships "sequence-affinity" of TBP binding to ssDNA [3] or dsDNA [4] yield the precise equilibrium equation:

$$-Ln(K_D) = 10.9 - 0.23 \times Ln(K_{D,2}) + 0.15 \times PWM - 0.2 \times Ln(K_{D,1}), \qquad (\#)$$

Results: This formula (#) predicts human diseases upon SNPs as it is exemplified by Table.

Table. The precise equilibrium equation of TBP/TATA-binding predicts human familial diseases upon SNPs

N	Allele	TATA-box sequence with **Mutation**	-ln(K_D)	δ	Health	Anamnesis	Disease
1	WT	acaggaccagCATAAAAggcagggca	18.94	0	Norm	Norm	Norm
	A-31**G**	acaggaccag**CG**TAAAAggcagggca	18.33	<	Illness	δ-Globin deficiency	δ-Thalassemia
2	WT	ttttgaaagcCATAAAAAcagcgagg	18.67	0	Norm	Norm	Norm
	C-31**T**	ttttgaaagc**T**ATAAAAAcagcgagg	19.85	>	Illness	Cell-cycle regulator excess	Lung cancer
3	WT	gccctcctgctatacagccccgccgc	18.80	0	Norm	Norm	Norm
	c-28**T**	gccctcctgcTATA**T**Agccccgccgc	19.64	>	Illness	Pseudogene act thumbs lens	Cataract
4	WT	gccag**G**TATAAAAAgggcccacaaga	17.36	0	Norm	Norm	Norm
	Δ-31**G**	ggccagTATAAAAAgggcccacaaga	17.23	<	Illness	Growth hormone deficiency	Short stature
5	WT	ccgggaatggAATAAAgggacgcggg	18.18	0	Norm	Norm	Norm
	g-34**T**	ccgg**T**aatggAATAAAgggacgcggg	18.25	>	Illness	Iron transporter excess	Female anemia

δ, SNP-caused TBP/TATA-affinity increase (>) and/or decrease (<). Genes: 1) δ-globin (GenBank: U01317) [Frischknecht & Dutly (2005) Hemoglobin **29**:151]; 2) interleukin 1β (AY137079) [Zienolddiny et al. (2004) IJC **109**:353]; 3) γEψ-pseudocrystallin (S72943) [Brakenhoff et al. (1994) HMG **3**:279]; 4) growth hormone-1 (J03071) [Horan et al. (2003) HUMU **21**:408]; 5) transferring (AJ252280) [Lee et al. (2001) BCMD **27**: 539].

where: the constant - nonspecific TBP/DNA-binding; $-Ln(K_{D,2})$ - sliding of TBP along DNA, as defined by [4]; PWM – TBP-sliding stop at native TATA-box, as given by [2]; $-Ln(K_{D,1})$ - endothermic transforms stabilizing TBP/TATA-complex, as denoted by [3].

Conclusion: This is the first attempt to calculate how TATA-box mutations alter human health.

Availability: Please email collaboration requests to Prof. N.A. Kolchanov, kol@bionet.nsc.ru.

References:

1. L. Savinkova et al. (2007) Interaction between the recombinant TATA-binding protein and the TATA-boxes of the mammalian gene promoters. *Ecological Genetics*. **5**: 44-49.
2. P. Bucher (1990) Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *JMB*, **212:** 563-578.
3. M. Ponomarenko et al. (1997) Modeling TATA-box sequences in eukaryotic genes. *Mol. Biol. (Mosk)*, **31**: 726-732.
4. M. Ponomarenko et al. (1999) Identification of sequence-dependent features correlating to activity of DNA sites interacting with proteins. *Bioinformatics*, **15**: 687-703.

TRNA'S FREE ENERGY AND EVOLUTION OF MITOCHONDRIAL GENOME

Popadin K.Yu.
Institute for Information Transmission Problems RAS, Bolshoi Karetny pereulok 19, Moscow 127994, Russia
e-mail: KonstantinPopadin@gmail.com

Motivation and Aim: Recently we created the manually curated database of secondary structures of all tRNA molecules from 277 tetrapod mitochondrial genomes [1]. Based on the structures we estimated Free Energy of each molecule and now we investigate relationships among stability of tRNA molecules and molecular evolution of the mitochondrial genes coding tRNAs (I) and proteins (II).

(I) We hypothesized that the rate of molecular evolution (rate of point substitutions as well as rate of structural changes in secondary structure) is slow down in stable tRNAs as compared with non-stable ones. Our preliminary tests confirm the hypothesis.

(II) Process of protein synthesis consists of two main steps. Firstly each amino acid is recognized by its cognate aminoacyl-tRNA synthetase (aaRS) and esterified to the specific tRNA to form an aminoacyl-tRNA (aa-tRNA). Secondly all aa-tRNAs are bound by elongation factor Tu·GTP (Ef-Tu·GTP) to form a ternary complex, which subsequently binds to the ribosome. On the both these steps structure of the tRNA molecules may play essential role in maintenance translation accuracy and efficiency. So we hypothesized that Free Energies of tRNA's may influence translation process through stability of intermediate products such as aaRS and ternary complex. We found out negative regressions among press of non-cognate amino acids and tRNAs' Free Energies as well as positive regressions among amino acid affinity and tRNAs' Free Energies, which is congruent with proposed hypotheses and argued that evolution of tRNA's molecules is dependent on amino acid composition of mitochondrial encoded proteins.

References:

1. K.Yu. Popadin, L.A. Mamirova, F.A. Kondrashov. (2007) A manually curated database of tetrapod mitochondrially encoded tRNA sequences and secondary structures. *BMC Bioinformatics*. **8**: 441

COMPUTER-AIDED PREDICTION OF BIOLOGICAL ACTIVITY SPECTRA FOR SUBSTANCES: VIRTUAL CHEMOGENOMICS

Poroikov V.V.*, Filimonov D.A., Gloriozova T.A., Lagunin A.A., Druzhilovsky D.S., Zakharov A.V., Stepanchikova A.V.
Institute of Biomedical Chemistry of Rus. Acad. Med. Sci., 119121, Moscow, Russia
* Corresponding author: vladimir.poroikov@ibmc.msk.ru

Motivation and Aim: The numbers of pharmacological targets and compounds available for screening have significantly arized in recent years making it impossible to test available libraries against all possible targets. Computational prediction of biological activity spectra for chemicals reduces the risk of missing useful pharmacological as well as unwanted adverse/toxic effects.

Methods and Algorithms: We have developed the PASS software that predicts about 3300 types of biological activity based on structural formula of drug-like organic molecules with average accuracy about 94% [1, 2]. Prediction is realized through the ligand-based design approach on the basis of structure-activity relationships established for about 120000 drugs, drug-candidates and pharmacological agents.

Results: PASS predicted the activity profile of ~250000 compounds from the NCI database with the hit rate increased up to ~17 times, and new pharmacological agents have been detected [3]. Application of PASS to selection of the most prospective compounds from virtual libraries significantly increases a probability of finding compounds with the required properties [4]. A dozen PASS applications were published in literature, where PASS predictions via Internet for diverse chemical classes and different types of activity were confirmed by further chemical synthesis and biological testing [5].

Conclusion: PASS can be used as *in silico* chemogenomics tool, for estimation of the most probable targets and biological effects arized from chemical compounds' action on these targets, thus significantly increasing the efficiency to find hits and leads with the required properties.

Availability: PASS INet predictions are available through http://www.ibmc.msk.ru/PASS

Acknowledgements: This research was supported in part by CRDF grant RC1-2064, RFBR grants 05-07-90123 & 06-03-08077, INTAS grant 03-55-5218, ISTC/BTEP grant 3197/111 and FP6 grant LSHB-CT-2007-037590.

References:

1. V. Poroikov, D. Filimonov (2005) PASS: Prediction of Biological Activity Spectra for Substances. In: *Predictive Toxicology*, C. Helma (Eds.), 459-478 (Taylor & Francis).
2. D.A. Filimonov, V.V. Poroikov (2006) Prediction of biological activity spectra for organic compounds, *Russian Chemical Journal*, **50:** 66-75.
3. V.V. Poroikov et al. (2003) PASS Biological Activity Spectrum Predictions in the Enhanced Open NCI Database Browser, *J. Chem. Inform. Comput. Sci.* **43:** 228-236.
4. A. Geronikaki et al. (2004) Design of new cognition enhancers: from computer prediction to synthesis and biological evaluation, *J. Med. Chem*., **47:** 2870-2876.
5. A. Geronikaki et al. (2008) Computer-aided predictions for medicinal chemistry via Internet, *SAR and QSAR in Environ. Res.*, **19:** 27-38.

LIFE WORKS ON AC POWER: THE IMPORTANCE AND PREVALENCE OF RHYTMS IN GENE EXPRESSION

Ptitsyn A.A.
Department of Microbiology, Immunology and Pathology, College of Veterinary and Biomedical Sciences, Colorado State University, Fort Collins, CO 80523
e-mail: Andrey.Ptitsyn@colostate.edu

Motivation and Aim: Periodic patterns in cellular processes are arguably the most underestimated factors in molecular biology. In 2006-2007 we have published a series of papers characterizing circadian oscillation in transcriptome of murine peripheral tissues, metabolic oscillation in yeast, circadian oscillation in plants.

Methods and Algorithms: We have implemented a panel of classic (Spectral analysis, fisher's g-test, autocorrelation, digital filters) and novel algorithms (Pt-test, K-S test for permuted periodogram, phase continuum, stochastic resonance) for the analysis of periodicity in gene expression. The data was taken from the public sources, contributed by the authors of previously published papers and partially produced in experiments with collaborators from Pennington Biomedical Research Center (Baton Rouge, Louisiana).

Results: Using a panel of advanced computational approaches we can demonstrate oscillation of a baseline expression in almost all genes interrogated by microarray as well as RT-PCR confirmation for a selected subset of genes previously considered non-oscillating. Only a small fraction of genes oscillate in the same phase in different tissues in mice. Circadian amplitude does not vary between tissues in most murine genes, but for some genes it is tissue-specific. The oscillatory pattern also changes between tissues and experimental conditions, which makes possible to analyze expression of very low expressed genes (previously considered "silent", i.e expressed below the resolution ability of either microarrays or RT-PCR) using the stochastic resonance approach. Among the most pronouncedly oscillating are "housekeeping" genes, which has to be taken in account in RT-PCR experiments. We also report a discovery of tissue-specific alternative transcription revealed by distinctive oscillation patterns. Alternatively polyadenylated transcripts with different turnover rate (determined by 3' UTR instability elements) can oscillate in counter-phase. Imbalance in alternative transcript population may cause dysregulation of cytokine signaling and interfere with the function of other biological pathways. There is a strong indication that circadian rhythms dominating mammalian gene expression are driven by the primordial oscillators of metabolic origin, synchronized to major environmental factors.

Conclusion: We have accumulated sufficient evidence to postulate that **a) all genes oscillate**; **b) genes can be expressed at a very low level, but never completely silent; c) frequency, phase and amplitude are important characteristics of gene function**. In the context of oscillating expression cell signaling causes not only up- or down-regulation and not switching the genes on or off, but a perturbation spreading waves through the biological pathways. Timing of the signal is an important factor as well as coordination of rhythms within a pathway. Not a single gene's function can be understood without putting it in a time prospective.

PREDICTION OF FUNCTIONALLY RELATED PROTEINS: PHYLOGENETIC PROFILES AND CLUSTER ANALYSIS

Pyatnitskiy M.A.*, Lisitsa A.V., Archakov A.I.
Institute of Biomedical Chemistry, RAMS, Moscow, Russia
e-mail: mpyat@mail.ru
* Corresponding author

Motivation and Aim: The advent of whole-genome sequencing has led to computational methods that infer protein function and linkages. One of the most promising approaches for prediction of protein-protein structural and functional interactions is studying of phylogenetic profiles [1-3]. A phylogenetic profile of a protein is a binary vector, representing the presence or absence of homologs to that protein across a set of organisms. It was shown that proteins with similar patterns of co-occurrence across many organisms tend to participate in the same protein complex, biochemical pathway or have similar sub-cellular location. In the present work we explored the application of cluster analysis to phylogenetic profiling in order to improve performance of the method.

Methods and Algorithms: We applied several standard techniques of cluster analysis including hierarchical clustering, kNN, PAM. We also proposed to use ART1 clustering, which is based on neural networks and was intentionally designed to handle binary vectors. This algorithm also has an advantage of automatically determining required number of clusters. KEGG database was used as a ground truth, metabolic pathways were considered as clusters. All software was implemented as a set of platform-independent Perl and R scripts. Computations were carried using cluster of 32 2xOpteron 2.6 GHz.

Results: We employed several measures for comparison of clusterings, including Rand index, silhouettes and etc. Null distributions for all indices were also computed to evaluate statistical significance of partition. Ward method and complete linkage clustering showed the best agreement with expert clustering, while single linkage showed the worst performance.

Conclusion: We showed that application of cluster analysis could improve performance of phylogenetic profiling. Also we proposed use of standardized measures (based on comparison of clusterings) to evaluate numerically methods for prediction of functionally related proteins.

Availability: Source code is freely available on request from authors.

References:

1. J.Wu et al. (2003) Identification of functional links between genes using phylogenetic profiles, *Bioinformatics,* **19**(12)**:** 1524-30.
2. E.S.Snitkin et al. (2006) Comparative assessment of performance and genome dependence among phylogenetic profiling methods, *BMC Bioinformatics,* **7:** 420.
3. D.Barker, M.Pagel (2005) Predicting functional gene links from phylogenetic-statistical analyses of whole genomes, *PLoS Comput Biol,* **1**(1)**:** e3.

GENETIC CONTROL OF HYPERTENSION IN ISIAH RATS

Pylnik T.O.*, Smolenskaya S.E., Markel A.L., Redina O.E.
Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
e-mail: Smallynx@ngs.ru
* Corresponding author

Motivation and Aim: The genetic control of the hypertension in the ISIAH (inherited stress-induced arterial hypertension) rats was studied. ISIAH strain was selected for increased response of systolic arterial blood pressure (ABP) to a mild emotional stress caused by 0.5 h restriction in a cylindrical wire-mesh cage. As a result of the selection, the ISIAH rats acquired the number of characteristic features concerned the hypertensive status.

Methods: The chromosome loci for several physiological traits (ABP at rest condition and under the stress, body weight, plasma level of corticosterone, behavior in the open field test) were detected by QTL method. The hybridization of the Atlas rat stress array, (BD Biosciences, Clontech, USA) helped to compare the expression profile of the 207 stress-related genes in brain, kidney and adrenals of the hypertensive ISIAH and normotensive WAG rats.

Results: According to the hybridizations the following genes may be considered as differentially expressed in ISIAH and WAG rats: Ribosomal protein S19 (frontal cortex), Superoxide dismutasa 1 (*Sod1*) (hypothalamus, medulla), Nucleoside diphosphate kinase (*Nme1*) (hypothalamus, medulla), Finkel-Biskis-Reilly murine sarcoma virus (*FBR-MuSV*) (medulla), putative c-Myc-responsive (*rcl*) (kidney and adrenals). *Rcl* is known as an antiapoptotic protein. The location of the *rcl* gene on Chr.9 corresponds to the QTL peak found for the adrenal weight. Its expression was more then 2.5 fold lower in ISIAH than in WAG rats both in kidney and adrenals. The expression of the *rcl* in frontal cortex and hypothalamus didn't differ in two rat strains. Since the different level of expression may result from the different structure of the regulatory regions, the GenBank sequences for the 5'UTR and 3'UTR *rcl* regions were analyzed. The full 5'UTR sequence contained only 20 nucleotides, so it may be considered as unknown. Two different sites of polyadenylation (3'UTRs of 34 bps and 266 bps long) being differently used in the different tissues were found in GenBank. 3'RACE (rapid amplification of cDNA ends) method was performed to detect the site of polyadenylation being used in rat liver. It was shown the both known sites are in use.

Conclusion: Rcl expression is directly stimulated by *c-Myc* [1]. Angiotensin II receptors (subtype AT1) may drive cell growth and replication in the cardiovascular system through the activation of the G protein, phospholipase C, diacylglycerol and inositol trisphosphate pathway, inducing the increase of certain protooncogenes *(c-fos, c-myc and c-jun*) expression [2]. The differential *rcl* expression found in ISIAH and WAG rats may play an important role in hypertension development in ISIAH rats.

References:

1. Lewis BC, Shim H, Li Q, Wu CS, Lee LA, Maity A, Dang CV. (1997) Identification of putative c-Myc-responsive genes: characterization of rcl, a novel growth-related gene. *Mol Cell Biol.,* **17(9)**: 4967-78.
2. Rosendorff C. (1996) The renin-angiotensin system and vascular hypertrophy. *J Am Coll Cardiol*. **28(4)**: 803-12

METHOD FOR COMPLEXITY REDUCTION AND MODEL COMPARISON WITH APPLICATION TO NFκB SIGNALLING

Radulescu O.[1], Zinovyev A.[2, 3]*, Lilienbaum A.[4]
[1] University of Rennes, CNRS UMR 6025, Rennes, France
[2] Institut Curie, INSERM U900 'Computational systems biology of cancer', Paris, France
[3] Institute of Computational Modeling, SB RAS, Krasnoyarsk
[4] CNRS URA 2115, Faculté de Médecine Pitié-Salpêtrière, Paris
e-mail: andrei.zinovyev@curie.fr
* Corresponding author

Motivation and Aim: There are many needs in developing methods of automatized simplification of systems biology models. Models often contain unnecessary complexity which conceals design principles, and renders their analysis difficult. Sensitivity studies, critical parameters and regulation loops identification become easier for reduced models. To find common patterns (model comparison) models should be simplified to a common level of complexity. Model reduction can be also used for robustness studies [1, 2].

Methods and Algorithms: We introduce a methodology allowing to reduce and to compare systems biology models. This is based on several reduction tools. The first tool is a combination of Clarke's graphical technique, averaging and idempotent algebra. The second tool is the Karhunen-Loève expansion (known also as principal component analysis). The nonlinear dimension of the system invariant manifold is estimated by a third method.

Results: We show how models of NFκB signaling can be simplified. From a complex model of NFκB pathway with 38 chemical species we produce a series of simpler models (with minimal model containing 6 species) that conserve as possible properties of the initial model. In this series there exists a model of NFκB pathway similar to the model proposed elsewhere [3]. We estimate the dynamical dimension of every model in this hierarchy and show that the dynamical dimension serves as a good indicator of real model complexity (number of truly independent degrees of freedom) and robustness. We suggest that the models of high structural complexity but low dynamical dimension should behave robustly with respect to their parameter variation.

Conclusion: For modeling biological process it is advantageous to analyze not a single model but a hierarchy of models of decreasing complexity. Models of lower complexity are easier to analyze, while the most complex model is easier to compare with real biological processes.

Availability: MATLAB implementation of the methods is available by request

References:

1. Radulescu, O., A. Gorban, S. Vakulenko and A. Zinovyev (2006). Hierarchies and modules in complex biological systems. *Proceedings of European conference on complex systems '06*. Oxford, UK.
2. Gorban A. and Radulescu O. Dynamical robustness of biological networks with hierarchical distribution of time scales (2007) *IET Syst Biol.* **1**(4):238-46.
3. Lipniacki, T. and al. (2004). Mathematical model of nf-*k*b regulatory module. *J Theor Biol* **228**, 195-215.

EMPIRICAL POTENTIALS FOR INTERACTION OF PROTEINS WITH WATER MOLECULES AND IONS

Rahmanov S.V.*, Makeev V.Y.
GosNIIGenetika Research Institute, Moscow, Russia
e-mail: sergeira@inbox.ru
* Corresponding author

Motivation and Aim: While interactions of proteins with solvent, including water molecules and different ions, is critical for protein structure stability and function, our understanding of these processes is still very limited. Experimental data on positions and specificity of the so-called single atom ligands often lacks specificity, and comprehensive modeling approaches are absent, despite the urgent need for these tools by experimental and structural biologists.

Methods and Algorithms: We have created a novel method for modeling protein interactions with water molecules and different ions, based on the analysis of the known three-dimensional structures of macromolecules {{http://www.biomedcentral.com/1472-6807/7/19}}.

Results: Using the new potentials, we were able to predict locations of structure-bound water molecules with accuracy not available previously. We demonstrate that whole-structure solvation modeling with these potentials allows successful protein recognition even without any consideration of the protein internal interaction. Prediction of ion binding sites and specificities in protein structures assists researchers in cases where experimental data is not available.

Conclusion: Empirical atom contact potentials for interaction of proteins with single atom ligands, such as ions, and water molecules, present a powerful and much needed tool for protein structure analysis and protein interaction modeling.

Availability: The service is freely available for academic use at the following URL: {{http://bioinform.genetika.ru/projects/hydration_potentials/index.html}}.

References:

1. Rakhmanov, S. V., V. J. Makeev (2007). "Atomic hydration potentials using a Monte Carlo Reference State (MCRS) for protein solvation modeling." BMC Struct Biol **7**: 19.

HETERODIMERIC CONSTRUCTS OF ANTI-THROMBIN APTAMERS AS MODEL BIORECOGNIZING ELEMENTS WITH ENHANCED AFFINITY FOR BIOSENSING

Rakhmetova S.Yu., Ivanov A.S.*, Radko S.P., Gnedenko O.V., Bodoev N.V., Veselovsky A.V., Shcherbinin D.S., Archakov A.I.
Institute of Biomedical Chemistry RAMS, Moscow, Russia
e-mail: ivanov@ibmh.msk.su
* Corresponding author

Motivation and Aim: A success in development of the biochip-based proteomics relies to a large extent on an availability of highly affine and selective biorecognizing elements. Aptamers represent a new class of affine reagents with a great potential for protein detection in a biochip format. Due to oligonucleotide nature and the synthetic way of production, aptamers are much more technological compared to antibodies. Additionally, aptamers lend themselves to a molecular design. Aptamers recognizing distinct sites on a protein target can easily be combined into a molecular construct allowing of simultaneous multiple interactions between the construct and the target protein. As may be expected, the multiplicity of interaction has to result in a significant reduction of the dissociation rate of the formed complexes. Thus, biorecognizing elements with enhanced affinity can be produced using aptamers as building blocks.

Methods and Algorithms: Two well-characterized anti-thrombin aptamers (primary aptamers) binding to different sites on a thrombin molecule were used to construct a model aptamer-based biorecognizing element. Aptamer motifs were merged into polynucleotide chains (heterodimeric constructs) via poly(dT) linkers of various length. Affinity of aptamers and their constructs was tested by the Biacore technology. Interaction of thrombin with aptamers was modeled by molecular dynamics using AMBER 8.0 suit.

Results: Apparent equilibrium dissociation constants of the heterodimeric constructs with linkers of 25 to 35 nt have been found to be at least an order of magnitude less than those of the primary aptamers. Study of kinetics of thrombin interactions with primary aptamers and the most affine heterodimeric construct revealed an increase of the association rate and a significantly reduced dissociation rate for complexes of thrombin with the aptamer construct. The enhanced affinity was observed in a wide range of potassium concentrations ranging from 0 to 145 mM. Molecular dynamics simulations showed that structure of protein-aptamer complexes was stable without potassium ions present. Simulation results agree with the experimental observation.

Conclusion: The enhancement of affinity observed for the heterodimeric constructs of anti-thrombin aptamers is consistent with the model of bidentate ligand binding. Aptamer heterodimeric constructs can present a new subclass of reagents with enhanced affinity for use in the biochip-based proteomics. Further development of computational models simulating interactions between aptamers and their targets may assist in a rational design of the heterodimeric constructs.

MALVAC: DATABASE OF MALARIAL VACCINE CANDIDATES

Ramachandran S.*, Gorai R.$, Ahmed S.$, Ansari F.A.$
G.N. Ramachandran Knowledge Centre for Genome Informatics, Institute of Genomics and Integrative Biology, Mall Road, Delhi 110 007, India
e-mail: ramu@igib.res.in
* Corresponding author; $ have contributed equally

Motivation and Aim: Malaria is a major killer disease. Annually more than 500 million cases are reported and over 1 million deaths occur. The sequencing of genomes of the Plasmodium species causing malaria, offers immense opportunities to aid in the development of new therapeutics and vaccine candidates through Bioinformatics tools and resources.

Methods and Algorithms: The starting point of MALVAC database is the collection of known vaccine candidates and a set of predicted vaccine candidates identified from the whole proteome sequences of Plasmodium species. These predicted vaccine candidates are the adhesins and adhesin-like proteins from Plasmodium species, *P. falciparum*, *P. vivax* and *P. yoelii*. Subsequently, these protein sequences were analysed through many publicly available algorithms to obtain Orthologs, Paralogs, BetaWraps, TargetP, TMHMM, SignalP, CDDSearch, BLAST with Human Ref. Proteins, T-cell epitopes, B-cell epitopes, Discotopes, and allergen predictions totalling to analysis through 16 algorithms. All these information were collected and organized against the ORFids of the protein sequences. These information are relevant from the view point of Reverse Vaccinology in facilitating decision making on the most probable choice for vaccine strategy.

Results: Detailed information on the patterning of the epitopes and other motifs of importance from the viewpoint of reverse vaccinology has been obtained on the most probable candidates for vaccine investigation on proteins from three major malarial species *P. falciparum*, *P. vivax* and *P. yoelii*. The results are displayed in convenient tabular format and a facility to export the entire data has been provided. The MALVAC database is hosted on a Web server.

Conclusion: A web server MALVAC for facilitation of the identification of probable vaccine candidates has been developed.

Availability: The MALVAC server can be accessed at http://malvac.igib.res.in/

STOCHASTIC DYNAMICS OF A SELF-REGULATORY GENE

Ramos A.F.*, Hornos J.E.M.
Institute of Physics of São Carlos, USP, São Carlos, Brasil
e-mail: ramos.a.f@gmail.com
* Corresponding author

Motivation and Aim: Variation in a genetically uniform cellular population under constant surrounding conditions had been observed early and is attributed to inherent stochasticity in gene expression, where a huge network of genes and proteins interact leading a cell to present an individual phenotype. While initial efforts dealt with stochasticity broadly and under a conceptual point of view, recent advances on experimental techniques permits a detailed quantitative analysis of fluctuations on gene circuits as well as their building blocks. Therefore, new theoretical endeavors are necessary for an accurate numerical description of such phenomena, and the raising of a mathematical based biological picture. Here, we present a stochastic model to a binary self-regulatory gene that is solved analytically for stationary and dynamical states, and present a Lorentz-like Lie symmetry. This completely solved model can be useful to understanding experiments dealing with an isolated gene.

Methods and Algorithms: We wrote the model in terms of coupled master equations for protein creation and death processes in a cell. The solution was obtained by application of the generating function technique. The steady state solutions and symmetries was obtained by considering the time derivative as null. The dynamical solutions was addressed by considering the eigenvalues and eigenfunctions of the operator acting on the generating functions;

Results: The stationary solution of the model is given in terms of confluent hypergeometric functions and the symmetry is a Lorentz-like Lie symmetry SO(2,1). The meaning of the invariant of the algebra is clear: it is the fundamental decaying rate of the system to equilibrium. The dynamical solutions are written in terms of the confluent Heun functions;

Conclusion: This model can be used as a fundamental module in a theoretical bottom-up approach of a gene network. It is stochastic, entirely solvable, and the symmetry brings to the field the mathematical machinery of group theory in the search for a composition rule between two or more genes.

References:

1. A.F.Ramos, J.E.M.Hornos (2007) Symmetry and stochastic gene regulation, Phys. Rev. Let. **99**, 108103.
2. A.F.Ramos, J.E.M.Hornos (2008) Exact dynamical solutions of a stochastic self-interctive gene, *in preparation.*

EXPERIMENTAL AND THEORETICAL ANALYSIS OF FATTY ACID RESPONSIVE GENE REGULATORY NETWORK IN YEAST

Ratushny A.V., Ramsey S.A., Roda O., Smith J.J., Aitchison J.D.*
Institute for Systems Biology, Seattle, WA, USA
e-mail: jaitchison@systemsbiology.org
* Corresponding author

Motivation and Aim: Peroxisomes are essential eukaryotic organelles responsible for beta-oxidation of fatty acids (FA) and sequestration of peroxides. Peroxisomal functions are linked to several central human health concerns including, developmental neuropathologies, aging and heart disease. Peroxisomes are dramatically induced by FA in yeast. Although several key molecules of the FA-responsive yeast transcriptional network and their core interactions are known, a systems-level comprehension of how this network controls dynamic FA-induced gene expression is lacking. Understanding the regulatory networks of FA-induced biogenesis of peroxisomes in a quantitative manner demands an iterative cycle of experimentation, model development, and simulation-based prediction of dynamic behavior.

Methods and Algorithms: Simulations of the model were performed using the Dizzy software program (http://magnet.systemsbiology.net/software/Dizzy/) and Matlab (Mathworks). Parameters of the model are taken from the literature or estimated from steady-state and time-course expression data.

Results: A mathematical model of the oleate-responsive gene regulatory network in yeast is developed. The model describes the kinetics of four core oleate-responsive transcription factor genes (*ADR1*, *PIP2*, *OAF1* and *OAF3*), as well as the expression of two oleate-inducible reporter genes (*POT1* and *CTA1*). Simulations of the model suggest that: (1) the Adr1p-driven feed-forward network motif reduces the steady-state variability of expression of target oleate-responsive genes combinatorially regulated by the factors; (2) the Oaf3p-driven inhibitory feed-forward loop modulates the dynamic response of target genes to a transiently varying oleate concentration.

Conclusion: Thorough investigation of the fatty acid response network has significant potential for fundamental understanding of gene regulatory network behavior, and more specifically the analogous response in mammalian systems. The interplay of Adr1p, Oaf1p, Pip2p and Oaf3p represents a "regulatory motif" that is highly overrepresented in the "yeast regulome" suggesting that our detailed analysis will be applicable to other regulatory subnetworks.

Availability: Available on request from the authors.

Acknowledgments: This work was supported by grants GM067228 and GMO76547 from the U.S. National Institutes of Health.

MATHEMATICAL MODELING OF GENETIC REGULATION OF PYRIMIDINE BIOSYNTHESIS IN ESCHERICHIA COLI

Ri M.T.[2]*, Khlebodarova T.M.[1], Likhoshvai V.A.[1, 2]
[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
[2] Novosibirsk State University, Novosibirsk, Russia
e-mail: rim@gorodok.net
[*] Corresponding author

Motivation and Aim: Development of an *in silico* cell as a computer resource for the modeling and analysis of intracellular processes is a topical problem of the systems biology and bioinformatics. Within this direction, it is necessary to develop mathematical models of the genetic regulation of cell metabolic pathways, in particular, the regulation of pyrimidine biosynthesis. It is of interest, because its decision permits us to control metabolism of the bacterial cell. The aim of this work was to analysis molecular-genetic mechanism of pyrimidine biosynthesis in cell *Escherichia coli* using mathematical modeling.

Methods and Algorithms: The gene network of pyrimidine biosynthesis was reconstructed using GeneNet system [1]. The method of generalized Hill functions [2] was used for create a base of elementary mathematical models (BEMM) of enzymatic reactions and genetic elements of this process. MGSmodeller was used for constructing the model of pyrimidine biosynthesis and perfoming in silico experiments [3].

Results: Reconstructed gene network of pyrimidine biosynthesis contains 108 objects (protein, mRNA, genetic elements and metabolits), which cohere in integrated system of 132 of elementary events (reaction, regulatory interaction). The base of elementary mathematical models contain 97 enzymatic reactions and genetic elements. parameters of elementary models were selected according to kinetic characteristics of subsistem's function that were measured experimentally. Basing on BEMM the mathematical model was developed, describing dynamics of pyrimidine nucleotide biosynthesis in cell *E. coli*. The computer analysis of the pyrimidine biosynthesis dynamic at different mutations and conditions were made.

Conclusion: Gene network reconstruction and creation of the mathematical models describing the functioning efficiency of enzymatic systems and expression regulation of the genes encoding the corresponding enzymes is a necessary initial stage in constructing the general kinetic model of pyrimidine biosynthesis gene network. This model will allow to predict dynamics of the processes going on in the system considered, to study their mechanisms, to detect key components of the gene network, and also to analyze the effects of mutations on its function state. The model will be an integral component of the computer resource an *in silico* cell under development.

References:

1. E.A. Ananko et al. (2005) GeneNet in 2005, Nucl. Acids Res., 33: D425-D427.
2. V.A. Likhoshvai, A.V.Ratushny (2007) Generalized hill function method for modeling molecular processes, JBCB, 5, 521-531.
3. F.V. Kazantsev et al., (2008), MGSmodeller – a computer system for reconstruction, calculation and analysis mathematical models of molecular genetic system, this issue.

THE GENES *Eps*, CONTROLLING ULTRA-EARLINESS OF WHEAT *TRITICUM AESTIVUM* L. THEIR EXPRESSION AND EVOLUTION

Rigin B.V.*, Koshkin V.A., Lam N.D., Matvienko I.I.
N.I. Vavilov All-Russian Research Institute of Plant Industry, St. Petersburg
e-mail: b.rigin@vir.nw.ru
* Corresponding author

Motivation and Aim: Development rate of soft wheat *T. aestivum* L. (shoots – spikes) in determined by the genes *Vern1-3* (in other data *Vrn4-5* as well), controlling vernalization response and growth habit, and different number of alleles *Ppd1-3*, controlling photoperiod response of wheat plants. These genetic mechanisms provides effective adaptation to wheat plants under different growth conditions. Analysis of genetic structure of wheat earliness and ultra-earliness in of great importance for both science and practice.

Methods and Algorithms: We have studied 304 earliness varieties and lines of *T. aestivum* from different ecology-geographic zones.

Results and Conclusion: The composition and combination of *Vrn1-3*, controlling growth habit for this set are typical for species *T. aestivum* on the whole. All earliness samples are non or weak sensitive to a short 12-hour´s day. The difference between earliness samples on rate of development can be controlled by the other genes, which are likely to be differed from Vrn and Ppd. Ultra-earliness samples of this set have been studied as well. In accordance with our experience, ultra-earliness samples are characterized by: 1) combination of strong dominant allele *Vrn1* and dominant alleles of other genes *Vrn*; 2) availability of dominant genes *Ppd*, ultra-earliness off plants at long day condition and practically neutral photoperiod response; 3) availability of gene *Eps* (earliness *per se*), which secures earliness to plants *per se*. We have created and study some ultra-earliness forms of soft wheat named Riko. These forms have dominant alleles *Ppd1-3* and are less sensitive to 12-hour´s day in comparison with wheat collection from VIR, besides they have dominant alleles *Vrn1-3*. Lines Rico Are likely to reflect a limit of possible earliness for soft wheat. On different length of daylight the ultra-earliness lines both Rico and Foton have a gene *Eps*, but genes *Eps* Rico and *Eps* Foton are non-allelic. It is very important to emphasize that gene *Eps* has a complete expression in a present of strong dominant alleles *Vrn* and *Ppd* and it is inherited independently from these genes *Vrn* and *Ppd*. Gene *Eps* promotes rapidly development of line Rico during the whole ontogeny. That is why Rico can be used in wheat breeding in various agricultural areas, especially in European and Siberian ones with short frostless period. The is not essential difference Rico´s earliness and the earless varieties of *Hordeum vulgare* L. from VIR collection. The gene *Eps* is assumed to be more ancient than both *Ppd* and *Vrn* genes. Angiospermae plants arose as spring forms at the middle of chalk period in tropical zone. At the early period gene *Eps* of *T. aestivum* was likely to have more compound structure and was responsible for duration of vegetation period, temperature and light reactions of the environment. During the evolution and range expansion of plant genetic systems controlling reaction of temperature factor (gene *Vrn*) and photoperiod response (gene *Ppd*) were sprang up. Various ecology promoted to appearance of polymorphism *T. aestivum* concerning.

SPLICE SITE VARIATIONS IN HUMAN DISORDERS

Roca X., Krainer A.R., Sachidanandam R.*
Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA
e-mail: ravi.cshl.work@gmail.com
* Corresponding author

Motivation and Aim: Many human diseases, including Fanconi anemia, hemophilia B, neurofibromatosis, and phenylketonuria, can be caused by 5'-splice-site (5'ss) mutations that are not predicted to disrupt splicing, according to position weight matrices. In this study we aim to understand this phenomenon by devising new measures of 5' splice site strength and identifying SNPs at 5'splice sites that may have an impact on human disorders.

Methods and Algorithms: By using comparative genomics, we identify pairwise dependencies between 5'ss nucleotides as a conserved feature of the entire set of 5'ss. These dependencies are also conserved in human-mouse pairs of orthologous 5'ss. Many disease-associated 5'ss mutations disrupt these dependencies, as can some human SNPs that appear to alter splicing. The consistency of the evidence across a variety of sources signifies the relevance of this approach. We use the pairwise dependencies identified above to identify 5'ss SNPs that may play a role in complex diseases.

Results: We have identified new measures of 5'splice site strength and identified human SNPs in 5'splice sites that probably have an impact on complex diseases.

Conclusion: This study highlights new measures of 5'ss strength and identifies 5'ss SNPs that might be relevant for population genetic analyses.

Availability: The 5'ss SNPs that we have identified are listed in a table in the reference [1].

References:

1. X. Roca et al. (2008). Features of 5'-splice-site efficiency derived from disease-causing mutations and comparative genomics., *Genome Res.,* **18(1):** 77-87.

THE MOLECULAR-EVOLUTIONARY BASIS FOR VAVILOV'S LAW OF HOMOLOGOUS SERIES

Rogozin I.B.[1, 2*], Glazko V.I.[3], Koonin E.V.[1]
[1] National Center for Biotechnology Information NLM, National Institutes of Health, Bethesda, MD 20894, USA;
[2]Institute of Cytology and Genetics, SB RAS, Novosibirsk 630090, Russia;
[3]Timiriazev's State Agricultural University, Moscow 127550, Russia
e-mail: rogozin@mail.nih.gov
* Corresponding author

Motivation and Aim: Rare genomic changes (RGCs) that are thought to comprise derived shared characters of individual clades are becoming an increasingly important class of markers in genome-wide phylogenetic studies [1,2]. Recently, we proposed a new type of RGCs designated RGC_CAMs (after Conserved Amino acids-Multiple substitutions) that were inferred using genome-wide identification of amino acid replacements that were: i) located in unambiguously aligned regions of orthologous genes, ii) shared by two or more taxa in positions that contain a different, conserved amino acid in a much broader range of taxa, and iii) require two or three nucleotide substitutions [3]. When applied to animal phylogeny, the RGC_CAM approach supported the coelomate clade that unites deuterostomes with arthropods as opposed to the ecdysozoan (molting animals) clade. However, a non-negligible level of homoplasy was detected [3,4].

Results and Conclusions: We provide a direct estimate of the level of homoplasy caused by parallel changes, one of the major classes of events leading to homoplasy, among the RGC_CAMs using 462 alignments of orthologous genes from 19 eukaryotic species. It is shown that the evidence in support of the Ecdysozoa clade, in large part, can be attributed to parallel changes. Parallel changes occur much more often in relatively recently diverged lineages than in those separated from their last common ancestor by longer time intervals of time. This pattern seems to provide the molecular-evolutionary underpinning of Vavilov's law of homologous series and is readily interpreted within the framework of the covarion model of molecular evolution.

References:

1. A. Rokas, P.W. Holland (2000) Rare genomic changes as a tool for phylogenetics. *Trends Ecol Evol*, 15: 454-459.
2. J.L. Boore (2006) The use of genome-level characters for phylogenetic reconstruction. *Trends Ecol Evol,* 21: 439-446.
3. I.B. Rogozin, Y.I. Wolf, L. Carmel, E.V. Koonin (2007) Ecdysozoan clade rejected by genome-wide analysis of rare amino acid replacements. *Mol Biol Evol*, 24: 1080-1090.
4. I.B. Rogozin, Y.I. Wolf, L. Carmel, E.V. Koonin (2007) Analysis of rare amino acid replacements supports the Coelomata clade. *Mol Biol Evol*, 24: 2594-2597.

PHYLOGENETIC STUDIES OF PROKARYOTIC XYLOSEISOMERASE

Rozanov A.S.*, Pintus S.S.
Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
e-mail: rozanov@bionet.nsc.ru
* Corresponding author

Philogenetic relationship of xyloseisomerase (XylA) sequences [5.3.1.5] was analyzed in this work. D-Xylose isomerase (D-xylose ketol-isomerase, EC 5.3.1.5) is an enzyme involved in pentose metabolism in various microorganisms. This enzyme catalyze reversible isomerization of D-xylose to D-xylulose. The sequences of enzyme are divided into two classes. These classes are distinguished by large deletion presented in the sequences of first class. The sequences from different protein's classes have ~ 15-20% of homology. XylA its never housekeeping gene and can have big and interesting history of the evolution.

Two philogenetic trees were created using Maximum Likelihood methods. First philogenetic tree was created for 16s rRNA, to reflect the philogenetic relationships among the prokaryotic genera. The second tree was arranged for xylA protein sequences. The sequences used or philogenetic trees were from the full genome data base of prokaryotic species. (www.ncbi.nlm.nih.gov/sutils/genom_table.cgi) The spread way of xylA gene in the bacteria domane was studied in this work

Results: - xylA gene rarely range in the proteobacteria domain. Gamma proteobacteria has the sizeable exclusion; one cluster of the genera Gamma ProteoEnteroBacteria has xylA gene in the genome. This gene have been introduced into genome by Horizontal Gene Transfer. xylA gene is the gene of xyloseisomerase class II. In other cases when xylA gene presented in genom of proteobacteria the gene was found in an individual genus, and has big similarity to the gene of the groups from domain Firmicutes. Such fact can be a marker of Horizontal Gene Transfer mo little time ego.

- In Firmicutes domain many genomes have xyloseisomerase gene. We can suppose that ancestor of this prokaryotic domain might have class II protein in its genome. However, in the evolution process this gene was subjected to elimination and substitution. The resulting philogenetic tree of xylA gene is not similar in full to that of 16s rRNA gene;

- the majority of Actinobacteria domain exhibit xylA gene of class I. The lack of the gene is the rare fact in genera as the existence of class II proteins in bifidobacterium genus, this gene could be transferred from lactobacillales (Firmicutes).One can suggest that ancestor of this prokaryotic domain might have class I protein in the genome.

- in case if this gene haven't been transffered and/or eliminated in the isolated branch of philogenetic tree of genera, the protein tree could posess the larger similarity to the 16s rRNA one.

- we can see that xylA gene have two origins in history of Bacteria: domains Actinobacteria and Firmicutes.

MULTIPLE ALIGNMENT BASED ON SPECIES TREE

Rubanov L.*, Seliverstov A., Lyubetsky V.
Institute for Information Transmission Problems RAS (Kharkevich institute), Moscow, Russia
e-mail: rubanov@iitp.ru
* Corresponding author

Motivation and Aim: Finding multiple alignment of given sequences, optimal in terms of a fixed score, is an NP-hard problem. Therefore, various heuristic algorithms were proposed to solve it. Some programs such as MultAl (by A. Mironov, unpublished) and Mafft do not use phylogeny data. Another programs like Clustal, build the alignment in combination with a tree (previously unknown). TreeAlign relies upon a given species tree, but yields to our algorithm in performance and effectiveness aspects. Thus we proceed with study of this problem.

Methods and Algorithms: Given a species tree, what we consider in a node is a sequence of nucleotide frequency distributions, i.e. a sequence of four-dimensional vectors. For the tree leaves, these vectors are given data, they have one unity component and others are zero. The proposed algorithm proceeds from the leaves to the root. Two distribution sequences for a pair of daughter nodes are aligned, and half-sum of distributions at each position of the alignment makes up a distribution sequence for the parent. The pair-wise alignment uses common gap penalty function, but bonuses and penalties for non-gap positions vary depending on the position: they are calculated e.g. as 1-r, where r is a fixed distance function of two distributions. We tried variants with r being scalar square of distribution difference as well as distances in L_1 and L_2 distribution spaces. When the root distribution sequence is built, the algorithm proceeds backward, deploying gaps from the root sequence into descendent ones up to the leaves. The set of sequences for all leaves with gaps inserted in that way give us the multiple alignment sought-for, the algorithm result. A gap in the above alignment is represented by the zero vector, therefore sequences built for the internal nodes may contain distributions with a sum of the components being less than 1. If the given sequences use the extended nucleotide nomenclature, the components change in accordance with known relative frequencies of individual nucleotides, e.g. of A and G that are represented by the letter R. If the individual frequencies are unknown, we assign equal values to them, so that their sum is 1.

Conclusion: A novel heuristic algorithm is proposed and implemented in a fast program which finds appropriate multiple alignment as justified by biological results. For instance, the processing of 16 sequences with length of 120–223 nt on Pentium-4 PC takes less than 0.5 s. The algorithm correctly aligns known promoters and protein binding sites of DNA/RNA in gene leader domains. Specifically, we used it to find the sites described in [1]. Part of those sites are confirmed experimentally; we consider such results as biological validation of the algorithm. The proposed algorithm works well for source sequences with significantly different lengths, and in presence of tandem duplications and uncertain sites as well. This algorithm substantially differs from that of known ones.

This study was supported by International Science and Technology Center (3807).

References:

1. A. Seliverstov, V. Lyubetsky (2008) On evolution of promoters in plastomes. This conference: BGRS'08.

DESIGN OF AN OLIGONUCLEOTIDE MICROARRAY FOR TYPING INFLUENZA VIRUS A

Ryabinin V.A.*, Sinyakov A.N.
Institute of Chemical Biology and Fundamental Medicine, SB RAS, Novosibirsk, Russia
e-mail: ryabinin@niboch.nsc.ru
* Corresponding author

Motivation and Aim: Hybridization microarrays are increasingly being explored for use in diagnostic applications. The most important stage in the design of microarray is the choice of oligonucleotide probes and is always a difficult task due to the huge amount of genomic sequences available in public databases. Here we report the approach to design a microarray that allows the subtype identification of all influenza A virus hemagglutinins and neuraminidases: H1 through H15 and N1 through N9, respectively.

Methods and Algorithms: The method is based on selection of oligonucleotide probes that make it possible to determine the hemagglutinin and neuraminidase subtypes by the hybridization of hemagglutinin and neuraminidase DNA on the microchip. The probes are selected in several stages. At the first stage, the peptides common for a certain hemagglutinin or neuraminidase subtype but absent in all the rest subtypes of the protein analyzed are selected. The next stage is the calculation of oligonucleotide probes based on the structures of peptides and DNA sequences corresponding to the protein analyzed. Then the oligonucleotide probes most representative for a particular hemagglutinin or neuraminidase subtypes are selected. At the final stage, the oligonucleotide probes with the necessary melting temperature are selected.

Results: The principles for selection of oligonucleotide probes to be used in hybridization microarray for typing influenza virus A were developed as well as the corresponding software for computing such probes. About 10000 hemagglutinin sequences and 7000 neuraminidase sequences of type A influenza virus were extracted from the GenBank database (up to January 2008). They were used to select 20–30 oligonucleotide probes for each hemagglutinin and neuraminidase subtypes. The probes were tested using the DNA sequences of hemagglutinin and neuraminidase (GenBank data of February–March 2008), which demonstrated that the designed microarray provided type A influenza virus serotyping with a high reliability.

Availability: Unavailable

The work was supported by ISTC, project # 3803

PRECISE POSTTRANSCRIPTIONAL EXPRESSION REGULATION

Saifitdinova A.F.*, Rubel A.A., Galkin A.P.
Saint-Petersburg branch of Vavilov Institute of General Genetics RAS; Saint-Petersburg State University, Saint-Petersburg, Russia
e-mail: saifitdinova@mail.ru
* Corresponding author

Motivation and Aim: For a substantial number of genes the absolute amount of protein in the cell is not strongly correlated to the amount of mRNA. These conclusions were based on simultaneous measurement of mRNA and protein at just a single time point [1]. Mainly, protein concentrations depend on the translation rate and the degradation rate. Here we report UTR specific precise mechanism of posttranscriptional regulation of particular protein production.

Methods: Experiments were carried out with *Saccharomyces cerevisiae* strain BY4742 (Invitrogen) and its isogenic derivates *hsp104Δ::KAN*^{r} and *tif4632Δ:: KAN*^{r}. Using molecular cloning methods we created reporter system with green fluorescent protein gene (*GFP*) under control of *CUP1* promoter. Transcription rates were estimated using reverse transcription real time PCR technique with fluorescent TaqMan probes to reporter gene *GFP* and reference gene *ADH1*. Amount of proteins were compared using immunoblotting followed with densitometry. Sequence analysis of 5'-noncoding untranslated regions (UTR) was performed.

Results: We compared the amounts of the *GFP* mRNAs in experiments with regular Hsp104 production, its overproduction and absence of Hsp104 and found that Hsp104 positively regulated the GFP expression controlled by the *CUP1* promoter without any change in the transcription rate of the corresponding mRNA. Similar results were reported for Hsp101, plant ortholog of yeast chaperone Hsp104 [2]. Hsp101 binds to the degenerate tandem repeat $(CAA)_n$ in the UTR of the tobacco mosaic virus mRNA and positively regulates its translation via interaction with the eukaryotic translation initiation factor 4F (eIF4F) complex. We have found that start codon of the *CUP1* promoter is preceded by $(CAAT)_4$ repeats which can serve as Hsp104-binding site. Moreover in *tif4632Δ* strain with imperfect eIF4F Hsp104 does not affect the rate of GFP translation.

Conclusion: Hsp104 is well known as a personage responsible for dissociation of aggregates of heatdamaged proteins and prion aggregates into smaller pieces or monomers as well as participation in degradation of alien proteins by interacting with other chaperones and proteins of the ubiquitin complex. In this research we have found new function of yeast chaperone Hsp104. Our results demonstrates participation of Hsp104 in posttranscriptional regulation of differential gene expression. Especially interesting that the rate of gene expression can depend on particular sequence in UTR of its promoter.

Acknowledgements: This work was supported by the Russian Foundation for Basic Research (project no. 07-04-00873), the Ministry of Education and Science of the Russian Federation (PHII.2.2.2.3.10047), and the US Civilian Research and Development Foundation (project no. BRHE Y4-B-12-04).

References:

1. R.Brockmann et al. (2007) Posttranscriptional Expression Regulation: What Determines Translation Rates?, *PLoS Computational Biology*, **3(3):** e57.
2. D.R.Gallie et al. (1987) The 5'-leader sequence of tobacco mosaic virus RNA enhances the expression of foreign gene transcripts *in vitro* and *in vivo*, *Nucleic Acids Res*, **15**: 3257–3273.

CANALIZATION OF GENE EXPRESSION IN THE *DROSOPHILA* BLASTODERM

Samsonova M.[1]*, Manu[2], Surkova S.[1], Reinitz J.[2]
[1] St. Petersburg State Polytechnical University, Russia
[2] Stony Brook University, Stony Brook, NY, U.S.A.
*e-mail: samson@spbcas.ru

Motivation and Aim: Developing embryos exhibit a robust capability to reduce phenotypic variations which occur naturally or as a result of experimental manipulation. This reduction in variation occurs by an epigenetic mechanism called canalization, a phenomenon which has resisted understanding because of a lack of necessary molecular data and of appropriate gene regulation models. In recent years, quantitative gene expression data have become available for the segment determination process in the *Drosophila* blastoderm. At the same time, predictive theoretical models for gene regulation have been developed. Together these advances make it possible to precisely characterize the epigenetic mechanism of canalization by means of dynamical systems theory supported by quantitative gene expression data.

Methods: Acquisition and processing of quantitative data was performed as previously described [1]. The numerical implementation of the gene circuit equations is as described [2] with the addition of time varying external inputs.

Results and discussion: We have detected that extensive variation in early segmentation gene expression patterns is markedly reduced by the time gastrulation begins [1], and that in the gap gene system this reduction in variation is the result of cross regulation. We demonstrate the validity of this explanation by showing that variation is increased in embryos doubly mutant for *Kr* and *kni,* disproving competing proposals that canalization is due to an undiscovered morphogen, or that it does not take place at all. We further show that canalization can be understood in terms of dynamical systems theory. In the anterior half of the embryo, variation reduction occurs because the system's qualitative dynamics are controlled by point attractors, but in the posterior variation reduction is governed by an attracting manifold. These results demonstrate that a complex multigenic phenomenon can be understood at a quantitative and predictive level by the application of dynamical systems theory.

Availability: All data are available from authors.

References:

1. S. Surkova, D. Kosman, K. Kozlov, Manu, E. Myasnikova, A. A. Samsonova, A. Spirov, C. E. Vanario-Alonso, M. Samsonova and J. Reinitz (2008). Characterization of the *Drosophila* segment determination morphome. *Developmental Biology*, **313(2)**: 844-862.
2. J. Jaeger, S. Surkova, M. Blagov, H. Janssens, D. Kosman, K. N. Kozlov, Manu, E. Myasnikova, C. E. Vanario-Alonso, M. Samsonova, D. H. Sharp, and J. Reinitz (2004). Dynamic control of positional information in the early *Drosophila* embryo. *Nature*, **430:**368-371.

METABOLIC NETWORK ANALYSIS OF NEUROBLASTOMA TUMOURS WITH GENE EXPRESSION DATA

Schramm G.[1, 2+], Gaarz A.[1+], Eils R.[1, 2], König R.[1, 2*]

[1] Department of Bioinformatics and Functional Genomics, Institute of Pharmacy and Molecular Biotechnology, Bioquant, Im Neuenheimer Feld 267, 69120 Heidelberg, Germany

[2] Department of Theoretical Bioinformatics, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 280, 69120 Heidelberg, Germany

[+]equally contributing

[*] Corresponding author, e-mail: r.koenig@dkfz.de

Motivation and Aim: Gene expression profiling with microarrays has produced data on a genomic scale for a large variety of different organisms and diseases, including our experimental investigations on neuroblastoma tumours. Such information needs to be funnelled into functionally meaningful patterns and applications. Besides this neuroblastoma tumours show a very heterogeneous clinical picture ranging from rapid growth with fatal outcome to spontaneous regression or differentiation into benign ganglioneuroma. Therefore, diagnosis and specific treatment is crucial and can be supported by understanding the molecular functionality of the tumour.

Results: We used a pattern recognition method we developed for the metabolic network of *E. coli* [1]. Basically, it defines discriminating gene expression patterns in the network by mapping the data onto two-dimensional adjacency matrices and applying wavelet transforms on these matrices. To reduce the sparseness of the matrices, we divided the network into smaller sections using the KEGG pathway maps. We evaluated all KEGG maps which served as our clusters in respect to their ability to discriminate neuroblastoma tumours of patients with favourable and unfavourable outcome. The most significant patterns were found for purine, glutamate, pyrimidine and one carbon pool metabolism, indicating increased nucleotide production for proliferation (purine and pyrimidine metabolism), and a switch in the glutamate metabolism.

Conclusions: Our findings may serve for more detailed experimental investigations, especially to treat the glutamate and one carbon pool metabolism.

References:

1. König R, et al: (2006) Discovering functional gene expression patterns in the metabolic network of Escherichia coli with wavelets transforms, *BMC Bioinformatics, 7:119.*

ON EVOLUTION OF PROMOTERS IN PLASTOMES

Seliverstov A.V.*, Lyubetsky V.A.
Institute for Information Transmission Problems RAS (Kharkevich institute), Moscow, Russia
e-mail: slvstv@iitp.ru
* Corresponding author

Motivation and Aim: Transcription initiation sites were experimentally determined for many plastome genes in *Arabidopsis thaliana* and few other species, [1-4]. We analyzed regions in plastomes of Streptophyta upstream genes to identify conservative promoters and study their evolution.

Methods and Algorithms: Original algorithm comprises two stages: (1) identification of sequences similar to the known promoters and (2) multiple alignment of fragments extending from the coding region up to a few nucleotides upstream the putative promoter. Alignments were computed with a program developed in co-operation with L. Rubanov. *Results*: The algorithm identified conservative promoters upstream gene *psbA* in almost all Streptophyta, and in *Bigelowiella natans* from Cercozoa. For this gene, algorithm produced negative answer for *Cycas taitungensis* and *Anthoceros formosae*. Conservative promoters were identified in almost all Spermatophyta upstream gene *psbB* and in almost all land plants and algae *Chaetosphaeridium globosum*, *Staurastrum punctulatum*, *Zygnema circumcarinatum* – upstream the *psbE* gene. For the last gene, it produced negative predictions in algae *Chara vulgaris*, *Chlorokybus atmophyticus* and *Mesostigma viride*. The algorithm predicted promoters upstream *psaA* not only in almost all land plants, but also in all studied algae from taxon Streptophyta, except *Chlorokybus* and *Mesostigma*. The algorithm predicted promoters upstream *rbcL* in almost all land plants and in only one alga, *Chara*. Notably, the predicted promoters and their 5'-untranslated regions are conservative, albeit the 5'-UTRs to a lesser extent.

Conclusion: Tests demonstrated high performance of the algorithm in finding plastome promoters. In particular, promoters were predicted in almost all Streptophyta. Conservative nature of 5'-UTRs may suggest presence of protein-binding sites in them. This was indeed shown for a number of plastid genomes (ref. [5]), where such binding sites are involved in regulation of translation initiation or mRNA processing.

This study was supported by International Science and Technology Center (3807).

References:

1. W. Gruissem, G. Zurawski (1985) Analysis of promoter regions for the spinach chloroplast *rbcL*, *atpB* and *psbA* genes, *The EMBO Journal* **4** (13A): 3375-3383.
2. A. Homann, G. Link (2003) DNA-binding and transcription characteristics of three cloned sigma factors from mustard (*Sinapis alba* L.) suggest overlapping and distinct roles in plastid gene expression, *Eur J Biochem* **270** (6): 1288-300.
3. P. Westhoff (1985) Transcription of the gene encoding the 51 kd chlorophyll a-apoprotein of the photosystem II reaction centre from spinach, *Molecular and General Genetics* **201** (1): 115-123.
4. M. Swiatecka-Hagenbruch et al (2007) High diversity of plastidial promoters in *Arabidopsis thaliana*, *Mol Genet Genomics*, **277**: 725-734.
5. A. Seliverstov, V. Lyubetsky (2006) Translation regulation of intron containing genes in chloroplasts, *J. Bioinform. Comput. Biol*.**4** (4): 783-790.

ANALYSIS OF CONTINUOUS 119737 BP STRETCH OF SUBTELOMERIC DNA ISOLATED FROM TRITICUM AESTIVUM BAC-LIBRARY

Sergeeva E.M.[1*], Adonina I.G.[1], Afonnikov A.D.[1], Chalhoub B.[2], Salina E.A.[1]
[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia;
[2]URGV-INRA, Evry, France
e-mail: sergeeva@bionet.nsc.ru
[*] Corresponding author

Motivation and Aim: We performed the detailed analysis of continuous stretch of *T. aestivum* subtelomeric DNA. The study of subtelomeric regions is of great importance because these regions are essential for genome stability and faithful chromosome replication. The large-insert genomic libraries provide us the possibility to analyze long contiguous stretches of DNA. To date many researches are focused on the large-insert clones of *Triticeae* carrying loci of known genes and there is no investigation aimed immediately to subtelomeric DNA.

Methods and Algorithms: DNA sequencing was conducted using random shotgun approach. To annotate the subtelomeric sequence we used consequently: identification of repetitive elements by BLAST algorithms against TREP (Triticeae Repeat database at GrainGenes), RepBase and TIGR (The Institute for Genomic Research) Gramineae repeat databases; and search of genes by integrating results of predictor programs GeneMark.hmm and FGENESH and BLAST algorithms against dbEST and SwissProt databases. The alignments were performed by ClustalW program; phylogenetic trees were constructed with neighbor-joining method by MEGA4 software package. The subtelomeric inherence of BAC-clone and its particular sequences was proved by *in situ* hybridization.

*Result*s: We isolated BAC-clone carrying subtelomeric DNA from *T. aestivum* cv. Renan genomic BAC-library using wheat subtelomeric marker Spelt52 as a probe. The subtelomeric BAC-clone 2050O8 of 119737 bp contains 8,2% (27 copies) of Spelt52; 6,9% of identified genes. Transposable elements (TE) account for 32,1% of 2050O8, with retroelements make up only for 7,8% clone length. DNA-transposons are predominant and account for 22,8% of entire BAC-clone. Among them, full-length CACTA transposon Caspar covers 11,667 bp (40,2% of DNA transposons) and encodes transposase and CTG-2 proteins. Based on the degree of sequence conservation and results of *in situ* hybridization with transposase and CTG-2 sequences as probes, we concluded that CACTA transposon Caspar tends to accumulate in distal regions of wheats and *Aegilops* and its divergence correlates with the cereal genomes evolution. Integrating the data of 2050O8 clone's *in situ* hybridization and its fragments BLAST search against mapped contigs in wheat, we managed to establish the location of 2050O8 on the end of 4BL chromosome.

Conclusion: We first made a detailed analysis of the continuous DNA sequence from wheat 4BL subtelomeric region. By the means of combined phylogenetic and cytogenetic approach it was shown that CACTA DNA-transposon Caspar_2050O8 is characteristic element for *Triticum* and *Aegilops* subtelomeres.

DETECTION OF NEW POTENTIALLY ACTIVE DRE SITES IN REGULATORY REGION OF HUMAN GENES ENCODING COMPONENTS OF Ah RECEPTOR CYTOSOLIC COMPLEX

Shamanina M.Y.[1]*, Oshchepkova E.A.[1], Oshchepkov D.Y.[1], Katokhin A.V.[1], Furman D.P.[1, 2], Tsyrlov I.B.[3], Mordvinov V.A.[1]

[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
[2] Novosibirsk State University, Novosibirsk, Russia
[3] XENOTOX Inc., Scarsdale, USA
e-mail: marinash@bionet.nsc.ru
* Corresponding author

Motivation and Aim: The aryl hydrocarbon receptor (Ahr), ligand-activated transcription factor, participates in a wide range of critical cellular events in response to endogenous signals or xenobiotic chemicals (1) modulating expression of numerous genes in various species and tissues (2). Unliganded Ahr exists as a heterotetrameric cytoplasmic complex composed of Ahr (a proper ligand binding subunit), the immunophilin-like Aip and a dimer of Hsp90s (3). Upon binding a ligand (L), typified by the most potent xenobiotic known so far, 2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD), Ahr/L in the complex translocates to the nuclei, dissociates from it and forms transcriptionally active complex with Arnt. This complex binds dioxin responsive elements (DREs) and thus modulates transcription of numerous target genes, including extensively studied *Cyp1A1* (4). Hence, it is important that Ahr levels and activity themselves be well controlled in target tissues (1). However, transcriptional regulation of Ahr cytosolic complex key members still remains to be established. Therefore, additional and new DREs found in the genes would shed a light on potential mechanisms of expression, stoichiometry of unliganded Ahr core complex, and its degradation vs biosynthesis dynamics.

Methods and Algorithms: We addressed the problem using computational approach named SITECON designed for recognition of potential transcription factor binding sites (TFBS). The tool is based on conservative conformational and physicochemical properties detected for set of experimentally proven TFBSs (5).

Results: The following number of new DREs in 5'-flanking region of human genes was detected: one in *AHR*, five in *AIP*, eight in *HSP90AA1*, and three in *HSP90AB1*. Also three reliable DREs were predicted by SITECON within human *AHR* gene downstream of the TSS (6), only one of those was previously identified as reliable (7). The most of newly and previously predicted DREs were found to be conservative in mammals (6).

Conclusion: Recognizing new DREs in the sequences of AHR cytosolic complex genes and detecting four unusually located DREs in *AHR* will allow us to develop human cell model of regulation and feedback regulation that seems to function in modulating expression of the AHR core complex genes.

Availability: SITECON: http://wwwmgs.bionet.nsc.ru/mgs/programs/**sitecon/.**

References:

1. P.A. Harper et al. (2006) *Biochem Pharmacol*, **72:** 267-279.
2. K. Sogawa, Y. Fujii-Kuriyama (1997) *J Biochem (Tokyo),* **122:** 1075-1079.
3. J.R. Petrulis, G.H. Perdew (2002) *Chem Biol Interact,* **141:** 25-40.
4. J.P. Whitlock (1999) *Annu Rev Pharmacol Toxicol*, **39:** 103-125.
5. D.Y. Oshchepkov et al. (2004) *Nucleic Acids Res,* **32:** W208-212.
6. E.A. Nedosekina et al. (2007) *Organohalogen Compounds,* **69:** 1889-1892.
7. Y.V. Sun et al. (2004) *Nucleic Acids Res,* **32:** 4512-4523.

CONTRASTING FEATURES OF SEX AND AUTOSOME CHROMOSOMAL EVOLUTION IN MALARIA MOSQUITOES

Sharakhov I.V.[1]*, Sharakhova M.V.[1], Xia A.[1], Tu Z.[2], Shouche Y.S.[3]
[1] Department of Entomology, Virginia Tech, Blacksburg, VA, USA
[2] Department of Biochemistry, Virginia Tech, Blacksburg, VA, USA
[3] National Centre for Cell Science, Ganeshkhind, Pune 411 007 India
e-mail: igor@vt.edu
* Corresponding author

Motivation and Aim: Chromosomal rearrangements are often associated with differential adaptations of malaria mosquitoes to various environments. Polymorphic inversions tend to cluster on 2R arm suggesting existence of hot spots for generating and/or maintaining rearrangements. Localization of hot and cold spots for rearrangements could be useful for identification of genes involved in ecological adaptations and biologically important gene clusters. The local adaptation model predicts parallelism between the extent of chromosomal polymorphism and evolutionary rates of inversion fixation. The aim of this work was to determine a possible correlation between distribution of polymorphic inversions and rates of fixed genome rearrangements in subgenus Cellia.

Methods and Algorithms: A physical genome map of *Anopheles stephensi* has been developed by mapping *A. stephensi*, *A. gambiae*, and *A. funestus* cDNA and BAC clones to polytene chromosomes. This 1 Mb resolution map was compared with physical maps of *A. funestus*, and *A. gambiae*. The gene order comparison has been performed using the Multiple Genome Rearrangements (MRG) {{http://www.cs.ucsd.edu/groups/bioinformatics/MGR}} and Sorting Permutation by Reversals and block-INterchanGes (SPRING) {{http://algorithm.cs.nthu.edu.tw/tools/SPRING/index.php}} programs.

Results: Analysis of inversions fixation rates showed significant differences among the chromosomal arms. The small and large blocks of the conserved gene order have been identified among *A. stephensi*, *A. funestus*, and *A. gambiae*. The smallest conserved blocks were found on 2R and X chromosome. 2R had also the highest density of polymorphic inversions. The largest conserved blocks (up to 6 Mb long) have been found in the chromosomal arms 3R and 2L of *A. gambiae.* Interestingly, these genomic regions are free from polymorphic inversions in the three species. Correlation coefficient for autosomes between *A. stephensi* and *A. gambiae* was 0.755 and between *A. funestus*, and *A. gambiae* was 0.832. In contrast, X chromosome has the highest rates of inversion fixation but does not have any polymorphic inversions in *A. stephensi*, *A. funestus*, or *A. gambiae.*

Conclusion: This study found that the synteny has been preserved at the whole arm level and that chromosomal elements evolve at different rates in subgenus Cellia. Parallelism was found between the extent of chromosomal polymorphism and evolutionary rates of inversion fixation for the autosomes but not the sex chromosome. The results suggest a major role of 2R inversions in adaptive evolution of malaria mosquitoes. A smaller effective population size and molecular features of sex chromosomes could be responsible for the high rate of inversion fixation on the X chromosome.

STRUCTURAL DYNAMICS OF HETEROCHROMATIN PATTERN IN EVOLUTION OF MALARIA MOSQUITOES

Sharakhova M.V.[1]*, Brusentsova I.V.[2], Tu Z.[3], Sharakhov I.V.[1]
[1] Department of Entomology, Virginia Tech, Blacksburg, VA, USA;
[2]Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia;
[3]Department of Biochemistry, Virginia Tech, Blacksburg, VA, USA
e-mail: msharakh@vt.edu
* Corresponding author

Motivation and Aim: Heterochromatin is a rapidly evolving part of the eukaryotic genome that initiates and accumulates the genetic differences between species. Anopheline mosquito polytene chromosomes represent an ideal model for studying evolution of the heterochromatin. The goal of this study was to identify the pattern of structural variations of the heterochromatin between two principal malaria vectors, *Anopheles gambiae* and *An. stephensi.*

Methods and Algorithms: Primary mouse antibodies C1A9 for Heterochromatin Protein 1 (HP1) of *Drosophila melanogaster* and ADL67.10 for *Drosophila* lamin Dm0 (Developmental Studies Hybridoma Bank, The University of Iowa) were used for imunostaining the *A. gambiae* and *A. stephensi* polytene chromosomes. The heterochromatic and euchromatic regions of the *A. gambiae* genome were analyzed for AT contents using a C program named ATCONTENT [1], percentage of matrix scaffold associated sequences were identified by SMARTest [2], and gene density was analyzed by Biomart {{http://www.biomart.org/ }}.

Results: Two different types, diffuse and condensed heterochromatin, have been identified in *A. gambiae* and *A. stephensi* chromosome based on morphology. Imunostaining of HP1 and lamin revealed an alternative pattern of their localization between the species. Both proteins were concentrated in centromeric areas of *A. gambiae* and in internal regions of *A. stephensi* chromosomes. No antibodies have been detected in some telomeric regions of both species and in condensed pericentromeric heterochromatin of *A. stephensi.* The total number of sites was 128/158 in *A. gambiae* and 266/268 for in *A. stephensi* chromosomes for HP1/lamin, respectively. The pattern of major invariable sites for both proteins within the species was identical. Bioinformatic analysis of the heterochromatin in *A. gambiae* genome revealed higher AT content and five times lower gene density then in euchromatin. Enrichment of matrix associated regions has been determined in all heterochromatic areas. Detailed analysis of tandem repeats and transposable elements content is in progress.

Conclusion: A. gambiae and *A. stephensi* have their specific structural patterns of the heterochromatin in polytene chromosomes. HP1 and lamin antibodies are primarily associated with diffuse heterochromatic structures in telomeric, centromeric and internal chromosome regions.

References:

1. Z. Tu (2001) Eight novel families of miniature inverted repeat transposable elements in the African malaria mosquito, *Anopheles gambiae*. *Proc Natl Acad Sci USA,* **13**: 1699-1704.
2. M. Frisch, et al. (2002) A new tool for the *in silico* prediction of matrix attachment regions in large genomic sequences. *Genome Research,* **12**: 349-354.

EXPRESSION ANALYSIS OF NF-κB-REGULATED GENES IN BREAST CANCER. META-ANALYSIS OF FIVE MICROARRAY DATA SETS

***Sharipov R.N.*[1, 2, 3*], *Kondrakhin Y.V.*[1, 3], *Kel A.E.*[4]**
[1] Institute of Systems Biology, Novosibirsk, Russia;
[2]Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia;
[3]Design Technological Institute of Digital Techniques, SB RAS, Novosibirsk, Russia;
[4] BIOBASE GmbH, Wolfenbuettel, Germany
* e-mail: shrus79@gmail.com

Motivation and Aim: Increased activation of transcription factor NF-κB was shown to play important part in human cancer development, besides other pathologies like autoimmune diseases, inflammation, and various viral infections. By blocking apoptosis NF-κB induces cell survival and proliferation. Elevated levels of NF-κB were revealed in many types of cancer (including breast cancer) and were also associated with phenomenon of chemotherapy resistance. In such way, NF-κB pathway is the subject of inquiry and targeting with various types of inhibitors.

Methods and Algorithms: Method IDURO (Identification of Down- and Up- Regulated Objects) developed by our group was applied to five independent sets of breast cancer cDNA microarray data obtained from the Stanford Microarray Database (http://genome-www5.stanford.edu) to reveal up- and down-regulated genes (in press). IDURO takes advantage of the hyper-geometrical distribution and is key for meta-analysis. Meta-analysis assigns the meta-score to each analyzed gene. The meta-score is defined up to sign as 10-base-logarithm transformation of optimal *p*-value. Analysis of the lists of genes up- and down-regulated significantly in breast cancer was performed with use of EXPLAIN 2.3 analysis system. BioUML workbench (http://www.biouml.org) was used for formal description of pathways of dysregulated genes and representation as diagrams.

Results: The gene lists generated with IDURO were analyzed focusing attention on genes containing experimentally confirmed or predicted NF-κB sites. Totally, 311 NF-κB-regulated genes were found of 391 revealed in the course of literature data analysis. Expression of 48 of them was significantly changed (28 – up-and 20 – down-regulated, |meta-score threshold| = 2.31; p-value <0.01). Expression of NF-κB subunits genes (RelA, RelB, etc) was not changed. Analysis of selected genes using EXPLAIN 2.3 system and Proteome BKL Disease™ database (BIOBASE GmbH) and Gene Ontology demonstrated that they comprise 12.53% of all diagnostic markers, 19.94% - therapeutic targets, 16.30% - associated with molecular mechanisms of breast cancer, and 22.73% - diagnostic markers of ductal subtype of breast cancer. Investigated genes are associated with development of 80 different types of tumors and 170 other human diseases. The strongest association was observed for breast cancer. Obtained data emphasize importance of NF-κB pathway for drug targeting. Formal description of networks of revealed genes was partly performed. Results of this work were deposited in two databases: BMOND (gene networks) and Cyclonet (gene lists). All collected data will be used in computational pharmacology for prediction of new targets to design more effective and safe anti-cancer drugs – inhibitors of NF-κB and its pathway.

Availability: All obtained data are public available in the Cyclonet (http://cyclonet.biouml.org) and the BMOND (http://bmond.biouml.org) databases.

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

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

Sharipov R.N. [1, 2, 3*], ***Yevshin I.S.*** [1, 2], ***Kolpakov A.F.*** [1, 3]
[1] Institute of Systems Biology, Novosibirsk, Russia;
[2] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia;
[3] Design Technological Institute of Digital Techniques, SB RAS, Novosibirsk, Russia
* e-mail: shrus79@gmail.com

Motivation and Aim: The BMOND (Biological MOdels aNd Diagrams, former Biopath) database was developed previously for reconstruction of complex biological systems from a huge amount of experimental data using method of formal description of biological systems. The main blocks of BMOND data inflows are eukaryotic cell cycle regulation and cancer (Cyclonet and Net2Drug project), NF-κB pathway and inflammation, nucleosomal regulation of gene expression, human essential hypertension, and iron metabolism and oxidative stress. BMOND is aimed at comprehensive description of biological pathways relating to these blocks, modeling/simulation and free access to deposited data via Internet.

Methods and Algorithms: BioUML technology (http://www.biouml.org) was applied for the formal description of structure and functioning of complex biological systems and processes on different logical levels represented in diagrams, as well as for import of a range of SBML (http://www.sbml.org) and CellML (http://www.cellml.org) models. Five diagram types of four levels of structuring: 1) *semantic network* (ontology, semi-structured data), 2) *pathway structure*, 3) *metabolic pathway* and *gene network*, and 4) *pathway simulation* (kinetic data: constants, equations, events, etc) – were used for data visualization. BeanExplorer Enterprise Edition (http://www.beanexplorer.com) was used for BMOND integration with free international biological databases and development of web interface for user access via Internet. The most part of collected data was obtained by method of manual annotation of literature.

Results: Now BMOND contains about 483 diagrams and models created on the base of 1533 articles, of which regulation of eukaryotic cell cycle and cancer (253 diagrams and 629 corresponding articles), NF-κB pathway and inflammation (92 diagrams and 278 articles), nucleosomal regulation of gene expression (45 diagrams and 434 articles), and essential hypertension (66 diagrams and 94 articles), and iron metabolism and oxidative stress (39 diagrams and 88 articles) represent the most part. The process of BMOND integration with other international databases: UniProt (proteins, http://www.uniprot.org), ChEBI (chemical substances, http://www.ebi.ac.uk/chebi), IntAct (protein interactions, http://www.ebi.ac.uk/intact), Gene Ontology (genes and their products classification, http://www.geneontology.org), BioModels (biomathematical models, http://www.ebi.ac.uk/biomodels), and EnsEMBL (genes, http://www.ensembl.org) has been going on. BMOND uses respective entities from these databases to build diagrams and models. This database is also integrated with Cyclonet (http://cyclonet.biouml.org) – our specialized database on cell cycle regulation and microarray data.

Availability: Biopath/BMOND database is available: online at http://bmond.biouml.org, and also as MySQL dump (by request) or set of text files (by request).

Acknowledgements: This work was supported by European Committee grant №037590 “Net2Drug” and interdisciplinary project № 46 of Siberian Branch of Russian Academy of Sciences.

INTEGRATED APPROACH FOR MODELLING PHYSIOLOGICAL, BIOMECHANICAL, AND MOLECULAR-GENETIC ASPECTS OF HUMAN CARDIOVASCULAR SYSTEM IN HEALTH AND ESSENTIAL HYPERTENSION

Sharipov R.N.[1, 2, 3], Yevshin I.S.[1, 2], Leonova T.I.[1, 3], Semisalov B.V.[1, 3], Biberdorf E.A.[4], Trakhinin Y.L.[4], Puzanov M.V.[1, 3], Blokhin A.M.[4], Markel A.L.[2], Ivanova L.N.[2], Kolpakov F.A.[1, 3*]

[1] Institute of Systems Biology, Novosibirsk, Russia;
[2]Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia;
[3]Design Technological Institute of Digital Techniques, SB RAS, Novosibirsk, Russia;
[4] Sobolev Institute of Mathematics, SB RAS, Novosibirsk, Russia
*e-mail: fedor@biouml.org

Motivation and Aim: Essential hypertension (EH) and induced pathologies of cardiovascular system (CVS) are leading causes of mortality in developed countries. Because EH is a complex multifactorial disease, delicate approaches for treatment and prophylaxis should be used. Until the present, there is no consensus of opinion, general approach or catalogue, which would describe this formally. The aims of the work were: 1) computational modelling tools improvement; 3) design of an integrated CVS model on the base of mathematical and biological models describing hemodynamics and blood pressure (BP) regulation in normal and hypertensive state; 4) discrimination of CVS model's key nodes critical for EH development with further verification in experiments.

Methods and Algorithms: BeanExplorer EE (http://www.beanexplorer.com) and BioUML (http://www.biouml.org) technologies were used for design/development of BMOND (Biological MOdels aNd Diagrams) database and modelling data annotation and integration. Models of blood flow [1] and blood circulation system [2] were used as the basis for CVS modelling. Methods of lines and orthogonal marching were used for calculations. Model of CVS [3] was used for the renal function modelling.

Results: Four biological blocks “Cardiovascular system”, “The kidney”, “Water-salt balance regulation” and “Neurohumoral regulation” describing BP regulation were created on the base of 51 manually created diagrams. Physico-mathematical block was also created on the base of one-dimensional model of blood flow circulation in 55 main arteries [1] using methods of lines and orthogonal marching. Numerical data for blood flow dynamics were obtained for all arteries and cross-sections examined. Using BioUML workbench import of mathematical models [2] and [3] to BMOND was started to analyze and use them as the basis for cardiac and renal functions modelling.

Availability: Materials of the biological part of this project are available at in BMOND at http://bmond.biouml.org.

Acknowledgements: This work was supported by integration and interdisciplinary grant №46 of Siberian Branch of Russian Academy of Sciences.

References:

1. D.N. Lamponi. One dimensional and multiscale models for blood flow circulation. Pour l'obtention du grade de docteur es sciences. EP, Lausanne, 2004.
2. F.Karaaslan et al. (2005) Long-term mathematical model involving renal sympathetic nerve activity, arterial pressure, and sodium excretion, *Ann Biomed Eng.,* **33**: 1607-1630.
3. A.P.Proshin, Y.V.Solodyannikov (2006) Mathematical Modeling of Blood Circulation System and Its Practical Application, *Automation and Remote Control,* **67**(2): 329–341.

DISBALANCE BETWEEN INNATE IMMUNITY RESPONSE AND ANTIOXIDANT DEFENCE IN BLOOD AND ASCITES: INTEGRATION OF EXPERIMENTAL AND MATHEMANTICAL MODELING

Shatalin Yu.V.[1], Naumov A.A.[1], Sukhomlin T.K.[1], Ermakov G.L.[1], Potselueva M.M.[1], Sharipov R.N.[2, 3, 4], Yevshin I.S.[2, 3], Kolpakov F.A.[2, 4*]

[1] Institute of Theoretical and Experimental Biophysics RAS, Pushchino, Russia;
[2] Institute of Systems Biology, Novosibirsk, Russia;
[3] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia;
[4] Design Technological Institute of Digital Techniques, SB RAS, Novosibirsk, Russia;
* e-mail: fedor@biouml.org

Motivation and Aim: Tumour cells proliferation is known to activate innate immunity thus resulting in oxidative stress and inflammation, which affect many physiological functions. Recently we have observed a reverse dynamics of ROS generating capacity of phagocytes in blood and ascites of rats during transplanted Zajdela hepatoma growth [1]. We hypothesised that the reverse dynamics of some physiological parameters may reflect a disbalance in homeostasis of the two physiological fluids. If so, specific non-invasive treatment able to correct the disbalance when applied specifically and at appropriate time could support the innate immunity efficacy and prevent tissue injury. The main goal of this work was to check the hypothesis both experimentally and theoretically, to detect key nodes of the subsystems (ascites, blood, innate immunity) and to develop efficient therapeutic strategy.

Methods and Algorithms: Experimental model: Zajdela hepatoma transplanted into peritoneal cavity of Wistar rats. Experiment design: simultaneous measurement of multiple biochemical parameters of blood and ascites. Recently elaborated biochemical, spectral and immune assays methods were applied to follow concentrations of individual antioxidants, ROS, metabolites and regulatory protein in dynamics. BioUML (http://www.biouml.org) workbench was applied for the formal description, compartmental modeling and simulation of investigated subsystems. The values of modeling parameters were either obtained from literature or calculated from obtained experimental data. BeanExplorer Enterprise Edition (http://www.beanexplorer.com) technology was used for web access to the data placed in the BMOND database (http://bmond.biouml.org).

Results: A comprehensive dynamic study of multiple parameters (cells, ROS, markers of oxidative stress, and antioxidants) disturbed in the investigated subsystems was performed. We revealed that oxidative stress induced enhancement of antioxidant defence (e.g., ceruloplasmin increase) in blood, but not in ascites. It seems, that ascites environment possess only non-specific antioxidant defence. The compartmental model was developed, and its key nodes were also detected. It takes into account the transport of fluids, proteins (e.g., transferrin, ceruloplasmin, and albumin), low molecular weight antioxidants and inflammatory signals between blood plasma and ascites, simulates interaction between oxidative stress and antioxidant defence in blood and ascites.

Availability: Modeling data are available in BMOND at http://bmond.biouml.org.

References:

1. M.M. Potselueva et al. (1999) Generation of reactive forms of oxygen by polymorphonuclear leukocytes during hepatoma growth in the peritoneal cavity of animals, *Tsitologiia*, **41**: 162-166.

COMPARATIVE PHYLOGENETIC ANALYSIS OF OPISTHORCHIID SPECIES BASED ON NUCLEAR AND MITOCHONDRIAL SEQUENCES

Shekhovtsov S.V.*[*1], Katokhin A.V.[1], Konkow S.[1], Yurlova N.I.[2], Serbina E.A.[2], Vodianitskaia S.N.[2], Fedorov K.P.[2], Besprozvannykh V.V.[3], Ohyama F.[4], Sithithaworn P.[5], Loktev V.B.[6], Mordvinov V.A.[1]

[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
[2] Institute of Taxonomy and Ecology of Animals, SB RAS, Novosibirsk, Russia
[3] Institute of Biology and Soil Science FEB RAS, Vladivostok, Russia
[4] Kawasaki Medical School, Kurashiki City, Japan
[5] Department of Parasitology, Khon Kaen University, Khon Kaen, Thailand
[6] State Research Center of Virology and Biotechnology "Vector", Koltsovo, Novosibirsk, Russia
e-mail: mailto:shekhovtsov@bionet.nsc.ru
* Corresponding author

Motivation and Aim: Species of the family Opisthorchiidae representing a important health problem are poorly identifiable due to limited molecular data. In this study, we revealed genetic variability of 4 species of this family using 3 nuclear markers (*ITS1* and *ITS2*, rRNA internal transcribed spacers 1 and 2; *pmI9*, 9th intron of the paramyosin gene) and the partial sequence of the mitochondrial gene cytochrome c oxidase 1 (*cox1*) and reconstructed molecular phylogeny based on the data.

Methods and Algorithms: Specimens of *Opisthorchis felineus* were collected throughout West Siberia, *Clonorchis sinensis* specimens were from the Russian Far East and Japan, specimens of *O.viverrini* were from Thailand and samples of *Metorchis bilis* were from Novosibirsk oblast. Primers were designed using sequences taken from GenBank or reported by other authors [1]. Phylogenetic trees were built using Mega v3.1 (for Neighbor-Joininig and Maximum Parsimony) and PAUP v4.10 (for Maximum Likelihood). Number of bootstrap replicates was 10000 for NJ and MP and 100 for ML.

Results: Low intraspecific variance for all investigated species must be mentioned. *Cox1* diversity was below 3%, while for trematode species from other families this value is much greater, up to 25% for *Schistosoma mansoni* [2]. Phylogenetic analysis of the three concatenated nuclear markers showed that *C.sinensis* and *O.viverrini* grouped together. Similar trees were obtained with the amino acid sequence of *pmI9*. This supports the results of morphological studies which suggested that the genera Clonorchis and Opisthorchis should be synonimize. *Cox1* presented a completely different topology with *M.bilis* being closer to *C.sinensis* and *O.felineus* than *O.viverrini* did.

Conclusion: By analysing nuclear sequences (*ITS1*, *ITS2* and *pmI9*) and mitochondrial sequences (*cox1*) we showed that for the species studied phylogenetics trees differed significantly, what implies that reliable conclusions about phylogenies can be made only by involving additional markers. As concerning *M.bilis* our analysis doesn't support its divergence into another subfamily. Further studies are necessary to clarify this question.

References:

1. J. Bowles et al. (1993) Nuclear and mitochondrial genetic markers highly conserved between Chinese and Philippine *Schistosoma japonicum*, *Acta Tropica*, **55**: 217-229.
2. J.A.T. Morgan et al. (2005) Origin and diversification of the human parasite Schistosoma mansoni, *Molecular Ecology*, **14**: 3889–3902.

CONDITIONS OF CORRECTNESS OF MODELLING OF NON-LINEAR AND REVERSIBLE MATRIX PROCESSES BY THE DELAY EQUATION

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

Motivation and Aim: Modeling of processes of linear DNA, RNA and protein molecules synthesis in gene networks with respect of biochemical reactions on matrix stages of elongation lead to super-large systems of ordinary differential equations. It stages a problem of finding more economic ways of gene networks modeling. Under assumption of non-reversibility and linearity of elementary intermediate stages of elongation the correctness of passage from system of ordinary differential equations to the delay equation has been proofed before (Likhoshvai et al., 2004). However, in nature systems reactions of elongation are non-linear and reversible. Aim of this work is to provide sufficient conditions to keep correct modeling of reversible and non-linear indeed processes of DNA, RNA and protein synthesis by one delay equation. It will let use simple models of molecules synthesis as parts of more complicated systems that describe processes in live cell.

Methods and Algorithms: The proximity of solutions of ordinary differential equations system and delay equation has been numerically checked in STEP software package. To study and analytically proof of proximity the technique of Laplace transformation operator has been applied.

Results: In case of linear model with reversibility and sinks

$$\begin{cases} dx_1/dt = f(x_n) - ((n-1)/\tau_1 + \omega)\cdot x_1 + (n-1)/\tau_2 \cdot x_2, \\ dx_i/dt = (n-1)/\tau_1 \cdot x_{i-1} - ((n-1)/\tau_1 + (n-1)/\tau_2 + \omega)\cdot x_i + (n-1)/\tau_2 \cdot x_{i+1}, i = 2,..,n-2, \\ dx_{n-1}/dt = (n-1)/\tau_1 \cdot x_{i-1} - ((n-1)/\tau_1 + (n-1)/\tau_2 + \omega)\cdot x_i, \\ dx_n/dt = (n-1)/\tau_1 \cdot x_{n-1} - \theta x_n, \end{cases}$$

where $\omega -$ sinks, $x_n -$ product of synthesis, the proximity to delay equation model has been proofed when number of intermediate reactions n is big enough and speed of direct process is higher than speed of reverse process ($\tau_1 < \tau_2$).

$$\max_{t<T}|x_n(t) - y(t)| \underset{n\to\infty}{\to} 0, \qquad \begin{cases} dy(t)/dt = e^{-\tau\omega} f(y(t-\tau)) - \theta y(t), t > \tau \\ y(t) = 0, t < \tau \end{cases}, \qquad where\ \tau = \frac{\tau_1\tau_2}{\tau_2 - \tau_1}, T > 0$$

In case of non-linear system when speed of passage through intermediate stages is described by $(n-1)/\tau \cdot x_i/(\beta + x_i^{1+\gamma}), where\ \beta > 0$, the similar result has been obtained if $\gamma > 0$.

Conclusion: Sufficient conditions of correctness of matrix processes modeling by delay equation have been obtained. It allows to significantly simplify the process of modeling of nature gene networks by replacing thousands of equations corresponding to intermediate stages with one delay. Theorems establish connection of micro and macro levels of substance synthesis process.

References:

1. V.A.Likhoshvai, et al. (2004) Modeling by a Delay Equation of Multi-Stage Synthesis of a Sample Without Bifurcation., *Sib. J. of Industrial mathematics,* **7(1):** 73-94.

IMAGING GENOMICS/GENETICS & TEMPORAL AND SPATIAL RESOLUTION IN BRAIN FUNCTION STUDIES

Shvarev Y.N. [1, 2]
[1] Institute of Cytology & Genetics, SB RAS, Novosibirsk, 630090, Russia; [2] Karolinska Institutet, SE-171 76 Stockholm, Sweden

Imaging genomics is a form of genetic association analysis of physiological responses in the brain during specific information processing registered by non-invasive *in vivo* techniques (Hariri & Weinberger, 2003). Since genes are directly involved in the development and function of brain regions subserving specific cognitive and emotional processes, functional polymorphism in genes may be strongly related to specific neuronal systems, thereby having a considerably more robust impact at the level of CNS characteristics than at the level of behavior. Gene polymorphism effects were demonstrated recently for interactions between brain morphology (hippocampus, prefrontal cortex, amygdala), functioning (BDNF, glutamate, serotonin) and particular behavioral patterns (schizophrenia, anxiety) with neuroimaging techniques (Bigos & Hariri, 2007).

In view of the recent advances in functional neuroimaging, the current status of non-invasive techniques applied for human and animal brain mapping could be reviewed by integrating hemodynamic and electrophysiological principles. There are several functional neuroimaging techniques based on hemodynamic principle which reflect the neuronal activation indirectly: functional magnetic resonance imaging (fMRI), positron emission tomography (PET) and single-photon emission computed tomography (SPECT). More frequently used electrophysiological techniques include electroencephalography (EEG), sensory evoked potentials (EP), and magnetoencephalography (MEG). The coupling between hemodynamic response and neuronal activity (neurovascular coupling) is under intensive investigation, and the current data suggest that the hemodynamic response significantly correlates to neuronal activity, especially synaptic activity, rather than to spiking. Each technique has its own characteristic features in terms of spatial and temporal resolution, and therefore it is important to apply appropriate technique combination for solving particular problems. For example, EEG has excellent temporal resolution, but to identify a neuronal generator its spatial resolution is not sufficient. On the other hand, neuronal generators can be localized by hemodynamic technique, but hemodynamic response develops considerably later than neuronal reaction *per se*. Thus, the information obtained from fMRI or PET could be used for estimating the generator source from EEG or MEG. Moreover, exciting data presented recently showed that neuronal electrical activity (organotypic rat brain cultures *in vitro* and neural current imaging in model experiments) can be detected with magnetic resonance directly (Petridou et al. 2006; Kraus et al., 2008).

Further advances in these technologies will promote the understanding of precise functional specialization and inter-areal coupling within the CNS, and will enable investigation of functional impact of genetic variation on behavior as well as on neurological and psychiatric disorders.

CHROMOVIRIDAE LTR RETROTRANSPOSONS FROM MOSSES (BRYOPHYTA)

Smyshlyaev G.*[*]*, Novikova O., Blinov A.
Institute of Cytology and Genetics? SB RAS, Novosibirsk, Russia
e-mail: _gera@gorodok.net
[*] Corresponding author

Motivation and aim: Retrotransposons are the major component of plant genomes. Chromoviridae LTR retrotransposons are widely distributed in eukaryotic organism. Four distinct clades of chromoviral retroelements are described for the vascular plants to date. At the same time almost nothing is known about LTR retrotransposons repertoire in bryophytes genomes. Some retrotransposons, which were lost by higher vascular plants, could be retained in mosses from common ancestor of higher plants. The present study was initiated to investigate chromoviridae diversity in mosses.

Methods and Algorithms: The *Physcomitrella patens* genomic sequences data were directly obtained from DOE Joint Genome Institute (http://genome.jgi-psf.org). We used UniPro GenomeBrowser software for chromoviral elements identification and analysis. Multiple DNA alignments were performed by ClustalW and edited manually. Phylogenetic analyses were performed using the Neighbor-Joining (NJ) method in MEGA 3.0 program. Statistical support for the tree was evaluated by bootstrapping. BLASTN and BLASTX were used to search the EST database to determine whether the elements are transcribed in the *Physcomitrella* genome.

Results: We have combined a search of chromoviruses in *Physcomitrella* genomic sequences and experimental investigation of diverse moss species. The analysis of whole *P. patens* genome detected new PpatensLTR retrotransposons which belong to the novel moss-specific clade of chromoviruses. Phylogenetic investigation clearly demonstrates that the newly described clade is more closely related to the Fungi/Metazoa group of chromoviral retrotransposons than to the clades which were described for other green plants (mostly for angiosperms). PCR screening and dot-blot hybridization of the 34 bryophytes showed the wide distribution of the newly identified clade. Moreover, three additional bryophyte-specific clades were identified, which were also classified as the “retained clades” of retrotransposons since none could be classified with Viridiplantae-wide distribution. BLASTN and BLASTX results for *P. patens* EST and RT-PCR results for other mosses suggest that almost all Chromoviridae elements are transcriptionally inactive or the level of their activity is undetectably low.

Conclusion and availability: Our data suggest that newly described clades appeared before a divergence of plants and Fungi/Metazoa groups and were lost by higher plants (gymnosperms and angiosperms). The addition of information about biodiversity, distribution and organization of the LTR retrotransposons, as one of the most abundant genomic components, is very important for further understanding genome structure and evolution of mosses.

A STUDY OF THE ASSOCIATION OF E148E AND IVS5 (+219) C/T POLYMORPHISMS IN THE DOPAMINE-β-HYDROXYLASE (DBH) GENE AND OPEN ANGLE GLAUCOMA

Soboleva D.E.*[1], Gubina M.A.[1], Kulikov I.V.[1], Konovalova N.A.[2], Konovalova O.S.[2], Romaschenko A.G.[1]

[1] Institute of Cytology and Genetics? SB RAS, Novosibirsk, Russia

[2] Tyumen State Medical Academy, Tyumen, Russia

e-mail: dinara2084@mail.ru

* Corresponding author

Motivation and Aim: Glaucoma is a widespread affection of the eyes leading to poor acuity of vision and blindness. Over 5 million people are blind. It is estimated that 14% of those blind are cases of glaucoma [1]. Glaucoma is sporadic and it has ethnic specific features. Glaucoma is a complex of disorders associated with visible alterations in the optic nerve head, abnormalities in the visual field and is frequently associated with raised intraocular pressure [2]. The aim of the present work was to study association of E148E and IVS5 (+219) c/t polymorphisms in the dopamine-β-hydroxylase (DBH) (Ac NC_000009) gene, which encodes enzyme of the dopamine to noradrenalin conversion with a strong vasoconstrictive effect. This leads to an increase in intraocular pressure and the development of glaucoma. This study was based on estimation of the contributory polymorphisms in the dopamine-β-hydroxylase (DBH) gene to disease development. The synonymous polymorphism E148E (exon 2, G444A) resides in donor splice site (5'ss) at the splice junction (position -1 in the 5'ss) and it probably affects selection of this 5'ss by U1RNP. The second mutation IVS5 (+219 c/t) is relatively far from 5'ss. The splicing regulatory protein can potentially interact with this nonconserved zone of the intron.

Methods and Algorithms: To identify associations of the chosen polymorphisms, we analyzed two patient samples with primary open angle glaucoma (POAG), 23 subjects, average age 73 years and a sample of patients with myopia and juvenile glaucoma, 30 subjects, average age 17 years. A group of residents of Novosibirsk served as control, 104 subjects, average age 46 years. The DNA samples were typed for two selected DBH gene polymorphisms using originally developed methods based on RFLP.

Results: The study for IVS5 (+219 c/t) in the DBH gene demonstrated significant differences of the myopia–juvenile glaucoma patients from the control sample (p=0.033) due to an increase of the frequency of the heterozygous variant (73.3% versus 39.4%). In the group of patients with juvenile glaucoma and myopia for the E148E polymorphism, there was a significant increase in the heterozygous variant compared with the control sample (p=0.004). The trend towards increase in the frequency of the heterozygous variant for E148E and IVS5 in group with POAG (65.2%) proved to be insignificant (p=0.255, p=0.067, respectively)

Conclusion: Thus, this study demonstrated that E148E (G444A) and IVS5 (5 intron +219 c/t) SNP in the DBH gene may have a possible effect on the development of OAG. The data obtained may be used for evaluation of individual and population risk of OAG and for development of a differentiated program of primary OAG prevention. *Acknowledgements:* Work was supported by the program “Dynamics of the gene resources of plant, animals and human” of the Russian Academy of Science.

References:

1. S.Resnikoff (2003) Prevention of blindness in the world: problems and approaches, *Materials of Russian Interregional Symposium:* 11-19.
2. J.S.Wolffsohn, A.L.Cochrane (1998) Low vision perspectives on glaucoma, *Clinical and experimental Optometry,* **81.6:** 280.

CELL VOLUME AND SODIUM CONTENT IN RAT KIDNEY COLLECTING DUCT PRINCIPAL CELLS DURING HYPOTONIC SHOCK

Solenov E.I.
Institute of Cytology and Genetics? SB RAS, Novosibirsk, Russia
e-mail: eugsol@bionet.nsc.ru

Motivation and Aim: Kidney epithelium in collecting duct contacts with hypotonic fluid which osmolality varies significantly and the cells continually expose to osmotic stress. To avoid damage and perform their functions principal cells require an effective cell volume regulation mechanism. Despite its importance, very little is known about cell-volume regulation in OMCD cells.

The purpose of this study was to investigate the time course of the volume-regulatory response and intracellular sodium concentration ($[Na^+]_i$) in the principal cells of rat kidney OMCD epithelia during acute swelling in hypotonic medium.

Methods: Hypotonic shock was created by PBS diluted with 50% of water. Changers in cell volume were measured with calcein quenching method. Intracellular sodium concentration was studied with fluorescence dye Sodium Green.

Results: Principal cells of micro dissected OMCD fragments swelled very fast. The characteristic time of swelling was 0.65±0.05 s, and the volume increased more than 60% (92.9±5.6 and 151.3±9.8 μm^3 control and peak volume correspondently, $p<0.01$). After cell volume reached the peak of swelling the RVD began without lack period. The characteristic time of volume decreasing to new steady state level was 8.9±1.1 s. After restoration of the medium osmolality to normotonic, cell volume stabilized on significantly low level in comparison with control level (71.4±6.1 μm^3, $p<0.05$). During the hypoosmotic shock $[Na^+]_i$ decreased from control level in isotonic PBS to the low level in hypoosmotic solution (27.7±1.4 and 5.8±0.23 mM, $p<0.01$).

Conclusion: Calculation of sodium content per cell, shown the significant sodium entry into the cells, which produced a peak correlated with the peak of cell volume caused by swelling. The conclusion is made that in our model of hypoosmotic shock swelling activates transporters with high permeability for Na+ that provides significant sodium flux into the cells.

DEVELOPMENT OF TEST-SYSTEMS FOR GENETICALLY MODIFIED CROP DIAGNOSTICS USING REAL-TIME PCR

Startsev V.A.*, Kulaeva O.A.
Saint-Petersburg State University, Department of Genetics and Breeding, Biological Research Institute, St. Petersburg, Russia
"GMO-Test" Ltd, St. Petersburg, Russia
e-mail: startsev@pochta.ru
* Corresponding author

Motivation and Aim: now the genetically modified plant database {{http://www.agbios.com}} concludes 130 GM plant lines, related to more then 20 species. More part of these lines is cultivated in planet fields. By the end of 2007 the volume of area under crops has reached 110 million hectares. Since 2008 in Russia 16 lines are allowed for using in food, and since 1th April 2008 the obligatory GMO containing food marking was imposed. So, an important duty is to develop efficient quality and quantity methods for GMO containing crop diagnostics.

Methods and Algorithms: we analyzed international and European patent bases: {{http://www.wipo.int}}, {{http://www.espacenet.com}}; we used subject articles {{http://www.ncbi.nlm.nih.gov}}. Based on this information we constructed basic GMO lines contigs using program VectorNTI 10.0. Because of the great number of ambiguous positions the contigs were checked by sequencing. After alignment we picked out universal areas of transgenic insertions. Then the primers and probes for multiplex reactions (with FAM, ROX и R6G dyes) were designed. In this investigation the IRRM (Belgium) reference material was used. Also we used plasmids with certain concentrations, constructed by our laboratory.

Results: test-systems detecting 35S promoter and NOS terminator were developed for GMO screening. In the capacity of reference genes the specified plant genes were used: *ADH* (corn), *LEC* (soybean), *GluS* (sugar beet), *Gos* (rice), *Sucr* (potato). We developed test systems for GMO lines: soybean – GTS-40-3-2, A2704-12, A 5547-127; corn – GA21, T-25, NK-603, MON-810, MON-863; Bt11, Bt176, sugar beet – Sugar beet Line77, and rice – LL62.

Conclusion: the base of transgenic sequences of the wide list of plants was made. In this investigation was developed 12 test-systems for diagnostics of the most widespread transgenic plants resolved in Russia, and 2 test-systems for GMO screening. All test-systems allow a qualitative GMO analysis by real-time PCR, a number of systems allow spending a quantitative estimation.

In view of our experience to date and the collected information authors develop new test-system and methods of the analysis (in view of needs of Russia), and already created sets allow to spend all-round monitoring of food stuffs and raw materials on GMO presence.

Supported with grant: START-2007 №4823p/7202 from 26.03.2007, {{http://www.fasie.ru}}.

SITEGA METHOD APPLICATION FOR GENOME WIDE PREDICTION OF P53 BINDING SITES

***Stepanenko I.L.*[1, 2]*, *Levitsky V.G.*[1, 2]**
[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
[2] Novosibirsk State University, Novosibirsk, Russia
e-mail: stepan@bionet.nsc.ru; levitsky@bionet.nsc.ru
* Corresponding author

Motivation and Aim: The p53 tumor suppressor is a master transcription factor of cellular response to stress that regulates the expression of genes of cell cycle arrest, apoptosis, DNA repair. The p53 binds two invert repeat consensus sequence RRRCWWGYYY separated by 0-13 base spacer. The identification of transcription target of p53 is a potential key to understanding function of p53 and its signalling pathways in tumorogenesis, but the recognition of sites with various spacers is a complex problem using position weight matrix algorithms.

Methods and Algorithms: Nucleotide sequences of 47 functional p53 binding sites (BS) were retrieved from TRRD database [1]. Than full set was aligned with respect to 1st and 2nd conservative motifs (cores) of p53 BS. Two separate SiteGA recognition models [2] were constructed for both cores. The superposition of these models allowed searching BS with variable spacers. The SiteGA method applied genetic algorithm to infer specific set of locally positioned dinucleotides (LPDs). Thresholds 0.8473 and 0.7809 were selected for SiteGA recognition functions of the 1st and 2nd cores. These thresholds implied equal false positive rate 8E-4 of the SiteGA models for the 1st and 2nd p53 BS cores. Sequences of 23570 human genes containing 2000 bases upstream of annotated transcription start were downloaded from RefSeq database {{http://hgdownload.cse.ucsc.edu/goldenPath/hg18/bigZips/}}. We considered gene as a potential p53 target if it upstream region contained a potential BS with any spacer from 0 to 21.

Results: Totally ~1200 genes (5% of analysed set) were identified as potential targets. Approximately 800 (67%) of genes had zero spacer. Using DAVID Bioinformatics tools {{http://david.abcc.ncifcrf.gov/home.jsp}} these genes were classified into apoptosis (35), transcription regulation (87), transport (126), cell metabolism (349) and others genes. We identify 78 genes with spacer 1 bp and 85 genes with 6 bp, although the longer spacers were also observed. In accordance with experimental data we did not find any target with 3 bp spacers. Thus, in comparison with other popular p53 BS search tools [3] we found a higher percentage of potential sites (33%) with nonzero spacer. Our implementation of optimized position weight matrix method [2] also allowed predicting substantially fewer portion of p53 BSs with nonzero spacers. Most probably this reflects fuzzy positioning of local nucleotide context peculiar to SiteGA method. As previously known genes we found CD82, ATF3, APAF1, NOXA, PTEN etc. Potential p53 BS with nonzero spacers allowed us to find some novel target genes; among them we may emphasize genes related with apoptosis, oxidative stress and receptor genes.

Conclusion: The SiteGA model was successively applied for large scale genome prediction of bipartite p53 BS. Notable portion of potential p53 BSs with nonzero spacers adds to reconstruction of p53 wide-genome network.

References:

1. N.A. Kolchanov et al. (2002) Transcription Regulatory Regions Database (TRRD): its status in 2002. *Nucl Acid Res*, **30:**312-317.
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. C.L.Wei et al. (2006) A global map of p53 transcription-factor binding sites in the human genome. *Cell,* **124**:207-219.

BMI-1 CONTROLS MANY TRANSCRIPTIONAL REGULATORS ESSENTIAL FOR CEREBELLUM DEVELOPMENT

Subkhankulova T.*, Zhang X., Leung C., Marino S.
Institute of Cell and Molecular Science, Barts and the London, Queen Mary School of Medicine and Dentistry, 4 Newark Street, London E1 2AT, e-mail: subkhankul@hotmail.com
* Corresponding author

Motivation: Polycomb group (PcG) proteins are epigenetic chromatin modifiers involved in heritable gene repression and maintenance of cell identity and proliferation. Bmi1 (B-cell-specific Moloney murine leukaemia *virus integration site 1)*, a core component of PRC1, was demonstrated to be a key cell cycle regulator. We previously showed that Bmi1 is strongly expressed in proliferating granule cell progenitors during cerebellar development and is involved in control of cell proliferation. This mechanism is based, at least partly, on transcriptional Cdkn12 locus encoding INK4a/ARF inhibitors of cell cycle. Here we set out to characterize more in depth the contribution of Bmi1 to Sonic Hedgehog (Shh) signaling during granule cell development.

Methods: Granule cells were isolated from mouse cerebellum (P7) and cultured in vitro. Affimetrix arrays and QPCR were performed using standard procedures. All statistical analysis and calculations were performed in R-language environment.

Results: Analysis of microarray data (mouse Affimetrix expression arrays) revealed approximately 5,000 genes differentially expressed across all four samples: wild type cells (WT), Bmi1 knockout (KO) cells, wild type cells treated with Shh (WTs) and knockout cells treated with Shh (KOs). As expected Shh treatment strongly shifted the expression levels of many signal transduction pathways in both knockout and wild cell types. However, the most significant difference in such shifts between wild type and Bmi1-knockout cells under Shh treatment was detected for genes implicated in (i) cell cycle and (ii) neuronal differentiation control. Thus, cell cycle inhibitor p21 encoded by Cdkn1a locus, was more then two-fold up-regulated in KO-cells, and even higher in KO cells treated with Shh. Chromatin ImmunoPrecipitation (ChIP) data revealed that Bmi-1-containning PRC1 may directly interact with non-coding region of Cdkn1a gene. Besides, we found that many transcription factors involved in neurogenesis and cell differentiation such as Foxa2, Robo3, Onecut2, E2f2, Mrg1, Lhx2, Trb63 as well as well known targets of Bmi-1, HOX-group of transcriptional regulators, were strongly up-regulated in granule progenitor cells knockout animals. QPCR confirmed up-regulation of these genes in KO mice, therefore suggesting that not only Cdkn12-dependent mechanisms are essential for Bmi1-mediated control of cell proliferation and differentiation during cerebellum development.

Conclusion: The results of the investigation revealed that many transcription factors involved in neurogenesis can be direct targets of Bmi-1, therefore there are likely to be other Bmi-1-dependent mechanisms of neurogenesis then known before. The transcription factor network mediated by Bmi-1 seems to play important role in cerebellum development and is largely to be explored.

QUANTITATIVE STUDY OF SEGMENTATION GENE EXPRESSION IN DROSOPHILA HOMOZYGOUS KR MUTANTS

Surkova S.*[1]*, Manu*[2]*
[1] St. Petersburg State Polytechnical University, Russia
[2] Stony Brook University, Stony Brook, NY, U.S.A.
* e-mail: sestr_sve@mail.ru

Motivation and Aim: Krüppel (Kr) gene plays one of the key morphogenetic roles in early development of *Drosophila*. This is a "gap" gene coding for transcription factor which is necessary for primary zygotic regulation of body segmentation. To elucidate the regulatory role of this gene it is important to analyze quantitative expression patterns of other segmentation genes in homozygous *Kr* mutants. Previous qualitative studies didn't give the precise information on positioning and levels of gene expression in these mutants. Our results may be further used to infer regulatory interactions within the segmentation gene network using mathematical model [1].

Methods: Acquisition and processing of quantitative data was performed as previously described [2].

Results and discussion: As was previously reported, during cleavage cycle 14A in *Drosophila Kr* mutants the posterior domain of *giant* (*gt)* and *even-skipped (eve)* stripe 7 are significantly shifted to the anterior relative to their position in wild-type embryos. We didn`t detect this difference in positions until 13 and 26 minutes from the beginning of cycle 14A for *gt* posterior domain and *eve* stripe 7 respectively. During the latter part of cycle 14A, these domains shift by 12 and 5% embryo length as compared with wild-type. As zygotic gap proteins appear at cycle 12 - 13, our results point to the existence of significant delay in the influence of the absence of Kr protein on the behavior of expression domains. This suggests that zygotic gap-gap cross-regulation does not play a role in the positioning of the segmentation gene expression domains at early times and comes into effect only during cycle 14A. We have also detected that by the end of cycle 14A in *Kr* mutants the positions of posterior *gt* and *knirps* (*kni)* domains coincide. This contradicts with previous observations that *gt* domain occupies the position of mutant *Kr* [3]. Moreover, position of posterior *gt* domain in *Kr;kni* double mutants is the same as in *Kr* single mutants. This suggests the existence of some regulator, other than *kni*, which prevents *gt* and *kni* domains from expanding further to the anterior.

Availability: All data are available from authors.

References:

1. J. Jaeger, S. Surkova, M. Blagov, H. Janssens, D. Kosman, K. N. Kozlov, Manu, E. Myasnikova, C. E. Vanario-Alonso, M. Samsonova, D. H. Sharp, and J. Reinitz (2004). Dynamic control of positional information in the early *Drosophila* embryo. *Nature*, **430:**368-371.
2. S. Surkova, D. Kosman, K. Kozlov, Manu, E. Myasnikova, A. A. Samsonova, A. Spirov, C. E. Vanario-Alonso, M. Samsonova and J. Reinitz (2008). Characterization of the *Drosophila* segment determination morphome. *Developmental Biology*, **313(2)**: 844-862.
3. E. Eldon and V. Pirrotta (1991). Interactions of the *Drosophila* gap gene *giant* with maternal and zygotic pattern-forming genes. *Development* **111**, 367-378.

DISTRIBUTION OF ACTIVE SITE STRUCTURAL ANALOGS IN ENZYME 3D-STRUCTURES: COMPUTER ANALYSIS

Teeys E.S.*, Ivanisenko V.A.
Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
Novosibirsk State University, Novosibirsk, Russia
e-mail: jack@gorodok.net
* Corresponding author

Motivation and Aim: The problem of definition of phylogenetic relations between enzymes with low rate of primary structure homology still isn't solved. It is of interest, because its decision permits us to detect conversion of one enzyme to another in the course of evolution. In this work structural analogs of active sites are considered as enzyme active sites turned off during process of evolution. The aim of this work was to find the special features of the distribution of active site structural analogs in enzyme 3D-structures.

Methods and Algorithms: In current research information about atom location in chosen enzyme was extracted from PDB database. This data was used by PDBSiteScan [1] program for search of active site structural analogs. Then the distance between enzyme active site and detected active site structural analogs was measured. Information about location of active site in chosen enzyme was taken from PDBSite [2]. The histogram of distance distribution was made. The values in every column were normalized to the number of amino acids.

Results: 30 enzymes from different taxonomic groups were analyzed. Distance histograms show accumulation of structural analogs close to active site in enzyme 3D structures. It was found significant differences between these histograms and random distribution. Interpretation of these results from the position of classical theory of evolution was suggested. It shows, that evolution goes by gradual substrate specificity variation of enzyme active sites. New active sites appear by inclusion some residues of old active site and residues near to it, because catalytic pocket is favorable to side chains of amino acid residues approaching and therefore to new active site formation. Finally, it leads to active site structural analogs clusterization.

Conclusion: There was shown that spatial distribution of active site structural analogs depends on enzyme active site location. The data about active site structural analogs can be useful for us to define location of active site in spatial structure of enzymes *in sito* and to detect phylogenetics relations between enzymes, which have low rate of primary structure homology.

References:

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

“PROMETHEUS” TOOLKIT FOR AGILE DEVELOPMENT OF BIOLOGICAL DATA STORING AND ACCESS SOFTWARE

Timonov V.S.[1, 2]***, Miginsky D.S.***[1, 2]
e-mail: vtimonov@bionet.nsc.ru
1 Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
2 Novosibirsk State University, Novosibirsk, Russia
* Corresponding author

Motivation and aim: The current rate of increasing the amount of data in biological field leads to great number of new databases being developed. The problem of current importance is unification and simplification of their development. This problem concerns storing data, their input and representation. The convenient solution will provide developers with the simple and agile technique to implement such databases and will provide user with uniformly-looking interfaces for editing, executing queries and navigating through different data sources

Methods: The “Prometheus” system is based upon mechanism of automatic engineering the end-user interface using its meta-description. This method is much simpler than GUI written on high-level languages from scratch. Each end-user interface is described by special patterns using certain XML format. The distinctive feature of this method is the unified interaction between GUI-components and with different data sources without any involvement from the developer side.

“Prometheus” could be considered as the toolkit for rapid creating of the visual interfaces to represent, edit and input various biological data in different sources of the information supported by system. Using the given approach significantly reduces the development efforts of the program by several times.

"Prometheus" includes a set of GUI components for representing the most common biological data types. Each component possesses the ready-to-use functionality to work with representing and editing of the associated data. The modular architecture of system allows making necessarily convenient expansion of capabilities of operating with data. “Prometheus” could be upgraded with relatively low cost to build up WEB user interface from present meta-description in view of creation additional functionality.

To work with different data sources there are data providers capable to be adapted under any model of data. They allow one to work not only with database sources but with the various files storing the structured biological information.

For example one needs to create a database of genes with specific attributes. For this purpose the following steps must be performed: (1) create the Java-classes model of genes; (2) annotate classes by the special description; (3) use one of the available data providers or create new one; (4) describe of the GUI by special patterns. After all the steps are done the development of the new ready-to-use application is complete.

Results: Initial approbation has been done with the TRRD [1] database. In brief terms, the user interfaces providing represent and filtration information have been developed for the further activity. The user is able to see the information on genes available in database and also the detailed description of each gene separately. Implementation is based upon JAVA, Eclipse Rich Client Platform [2] and Hibernate [3] technologies that has provided use of the given development on various operation systems.

References:

1. TRRD database, http://wwwmgs.bionet.nsc.ru/mgs/gnw/trrd/
2. “Rich Client Platform”, http://wiki.eclipse.org/index.php/Rich_Client_Platform
3. Hibernate technology, http://www.hibernate.org/

NUCLEOTIDE ASSYMETRY IN STRUCTURAL RNAS: EVIDENCE OF C→U DIRECTINAL CHANGE IN TRNAS

Titov I.I. [1, 2]*
[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
[2] Novosibirsk State University, Novosibirsk, Russia
e-mail: titov@bionet.nsc.ru
* Corresponding author

Motivation and Aim: RNA nucleotide composition appears to be a highly variable character and, thus, can evolve. However, no universal trends in ongoing changes of structural RNA composition have been reported unlike of DNA or trends in amino acid frequencies of proteins [1].

Methods and Algorithms: For nucleotide frequency analysis I used mature miRNAs and miRNA genes and aligned sets of tRNAs, 5S RNAs, 16S RNAs and uRNAs. 20x123=2460 tRNAs, each corresponding one amino acid of one organism (123 organizms representing all three domains of life - Bacteria, Archaea and Eukaryota), were analyzed to reveal favored nucleotide substitutions by closest triples analysis similar to [1]. The fluxes of reciprocal substitutions were calculated for each of 20 tRNA families.

Results: All types of structural RNA exhibit the same pattern of nucleotide frequency correlations between G-, C- and U-nucleotides (while A correlates weakly with the rest): the more positions are involved into double-stranded regions of RNA secondary structure, the higher is the correlation within the RNA type.

In tRNAs G- and C-nucleotides were found to be most conservative. Transitions were most frequent amongst nucleotide substitutions. Directional change of C→U was revealed.

Ser-family was detected as most variable amongst 20 iso-functional tRNA families; no correlation between family variability and amino acid “age” was found.

Conclusion: tRNAs are not in detailed evolutionary equilibrium, consistently losing C- and accumulating U-nucleotides. This may reflect either under-representation of U in early tRNAs or, more probably, relaxed selection constraint favoring G-U pairs compared to G-C pairs in helical regions of secondary structure.

Acknowledgment: The Project "Evolution of molecular-genetic systems: computer analysis and modeling" of the RAS Presidium program "Biosphere origin and evolution" and the Project "System biology: computer-experimental approaches" of the RAS Presidium program of molecular and cell biology supported this work.

References:

1. I.K. Jordan et al (2005) A universal trend of amino acid gain and loss in protein evolution, *Nature,* **433:** 633-638.

COTRASIF: CONSERVATION-AIDED TRANSCRIPTION FACTOR BINDING SITE FINDER

Tokovenko B.T.[1]*, Golda R.Ya.[2]
[1] Institute of Molecular Biology and Genetics of NASU, Kyiv, Ukraine
[2] National University of "Kyiv-Mohyla Academy", Kyiv, Ukraine
e-mail: cotrasif@biomed.org.ua
* Corresponding author

Summary: A new tool has been developed for the genome-wide identification with increased specificity of the putative transcription factor binding sites (TFBS) in eukaryotic gene promoters.

Motivation and aim: Promoter analysis and TFBS identification are essential for the identification of gene regulatory networks. Low specificity of the TFBS prediction in eukaryotic gene promoters is a challenging task for modern bioinformatics.

Based on our previous research [1], we observed better specificity of the TFBS search when comparing the promoters of orthologous genes of the evolutionary close species (e.g. rat and mouse) for the presence of the target TFBS.

Our aim was to develop an easy-to-use web-tool for genome-wide identification of putative TFBS with enhanced results quality.

Methods and algorithms: COTRASIF is built upon the semi-automatic importer of promoters from the Ensembl automatic genome annotation database. Currently COTRASIF has 11 genomes available (including the popular human, rat, and mouse genomes).

Promoters are defined as 800bp upstream from transcription start site, plus the 5' UTR. For the initial TFBS search, either classical position-weight matrix (PWM) approach is used, or the recently developed HMM-based (hidden Markov models) search method. For PWM method, frequency matrices are needed as input; for HMM – a list of at least 3 known sequences of the TFBS, plus an optional position frequency matrix.

Initial search results can be further analyzed using the built-in gene orthology filter. Orthology information is automatically obtained from the Ensembl Compara genome alignments database. If the putative TFBS is present in the promoters of the genes of both orthologous genes being analyzed, then it has higher probability of being functional (biologically meaningful).

Results: We developed a web-accessible tool (conservation-aided transcription factor binding site finder, COTRASIF) for the genome-wide conservation-aided TFBS search.

Further development includes: addition of new genomes; integration of the Gene Ontology category enrichment functional analysis (hypergeometric and Bayesian); more convenient results output formats; specialized API for working with COTRASIF.

Availability: COTRASIF is freely available at http://biomed.org.ua/COTRASIF/

References:

1. B.T.Tokovenko et al. (2007) In silico approach to study and functionally analyze interferon regulated genes, *Biopolymers and cell*, **23**: 368-375.

IN SILICO PREDICTION AND FUNCTIONAL ANALYSIS OF PRIMARY INTERFERON-RESPONSE GENES

Tokovenko B.T.*, Obolenskaya M.Yu.
Insitute of Molecular Biology and Genetics of NASU, Kyiv, Ukraine
e-mail: b.t.tokovenko@imbg.org.ua
* Corresponding author

Motivation and aim: Our aim was to identify and verify the list of the genes of primary interferon response. Finding genes which have biologically meaningful ISRE (interferon-stimulated response element) is important for better understanding of the Jak-STAT activated cellular IFN response.

Methods: We used transcription factor binding site (TFBS) search with gene orthology filtering to find putative ISREs in the promoters of protein-coding genes of *Rattus norvegicus*, and used Gene Ontology (GO) analysis to check the validity of ISRE search results in terms of biological meaning. A total of 23286 promoters of rat genes were analyzed. To filter biologically insignificant results, we looked for ISRE occurrence in the promoters of orthologues rat and mouse genes, with no more than 25bp distance between them relative to the TSS (transcription start site) of each gene. GO enrichment analysis was carried out using BayGO tool for R environment.

In order to define the threshold to make a presence/absence call for each matrix-site similarity score, we obtained means and standard deviations of the maximal similarity score distributions for all genome's 2nd exons and all promoters. For exons, mean = 0.667, SD = 0.068; for promoters, mean = 0.767, SD = 0.043. Assuming all matches of ISRE in exons to be false-positive, threshold was chosen at the similarity score level of 0.80, which includes no more than 2.5% of high-scoring ISRE matches in exons, and includes 16% of promoter ISRE matches.

Results: Running TFBS search on both DNA strands with 80% threshold in the promoters of all the protein-coding rat genes produced 5 214 binding sites in 4 571 promoters. Of these, 850 ISREs in 768 promoters passed orthology-based selection.

Distribution of ISREs along the promoter in 768-gene set reveals 3 regions of ISRE localization: 0 to -250, -250 to -550, and above -550 relative to TSS. It is not yet known whether ISRE localization has any functional implications.

GO analysis of 768 gene set vs. all rat protein-coding genes produced 48 enriched GO categories ($p < 0.01$), with 13 categories related to known IFN effects. GO analysis of 768 gene set versus 4571 gene set produced 28 categories, with 9 related to known IFN effects. "Cell differentiation" and "development" in biological_process, as well as "binding" in molecular function had the lowest p-values in both GO analyses.

Conclusion: We identified 768 rat genes which contain ISRE in their promoters, and are the potential targets of transcriptional regulation by type I interferons. Functional analysis of these genes, conducted using Gene Ontology, had shown the relative enrichment of some of the GO categories related to already known IFN effects.

We identified 3 regions of dominating ISRE localization in promoters. Differences in functional roles of ISRE in these regions were not studied.

It was shown that additional orthology-based processing of the TFBS search results produces genes in those GO categories, which remain the most significantly enriched when comparing to all-genome protein-coding genes. This indicates that orthology-based selection helps to obtain more biologically significant search results.

Genes identified in this research as containing ISRE in promoters will be used to seed the construction of the IFN-α-induced gene regulatory network model.

AN INFORMATION ENTROPY MODEL FOR THE PHYLOGENESIS OF THE 1918 INFLUENZA VIRUS

Torrens F.*, Castellano G.
Institut Universitari de Ciència Molecular, Universitat de València, Edifici d'Instituts de Paterna, P. O. Box 22085, 46071, and Instituto Universitario de Medio Ambiente y Ciencias Marinas, Universidad Católica de Valencia *San Vicente Mártir*, 46003, València, Spain
e-mail: Francisco.Torrens@uv.es
* Corresponding author

Motivation and Aim: The amino-acid compositions of certain avian lysozymes was determined. The amino-acid sequence of hen egg-white lysozyme enzyme was annotated. Certain discrepancies exist between this sequence and others. Crystallographic analysis gave some results that are in agreement with the former. Discrepancies at residues 46, 48, 65 and 66 are a difference between Asp or Asn. Comparative studies are interesting from the viewpoint of structure–function relationships.

Methods and Algorithms: Differences between avian species sequences are expressed as percentage of different amino acids in lysozyme. The greater the differences, the farther in time must be separation between species. Optimality criterion SS associated with different proposals for phylogenetic trees allows *equipartition conjecture* to be validated/invalidated in phylogenesis. If, in the calculation of *entropy* associated with the phylogenetic tree, a species is systematically omitted, difference between entropy with and without this species can be considered as a measure of *species entropy*.

Results: The Trp-62 of hen lysozyme, which also plays an important role in substrate binding, is replaced by Tyr in human lysozyme. *Grouping level b* can be identified with *biological time*. Obtained *phylogenetic tree* is represented. The scheme is in agreement with data obtained in morphological studies. In the search of the keys of the origin of 1918 virus hemagglutinin (HA), the gene sequences of HA subtype H1 of several strands of influence virus were analyzed. Its phylogenetic tree was built.

Conclusion: It is not within the scope of the simulation method to replace biological tests of drugs or field data in palaeontology, but such simulation methods can be useful to assert priorities in detailed experimental research. Available experimental and field data should be examined by different classification algorithms to reveal possible features of real biological significance. The samples of 1918 strand are inscribed in that family of influenza virus adapted to man. Distance between 1918 gene H1 and known avian family reflects that it was originated in a strand of avian influenza virus, although evolved in unidentified host before emerging in 1918.

Availability: Available on request from the authors for academics; available as a commercial package for enterprises.

References:

1. F. Torrens, G. Castellano (in press) Periodic classification of human immunodeficiency virus inhibitors, In: *Biomedical Data and Applications*, A. S. Sidhu, T. S. Dillon, E. Chang (Eds.), Springer, Berlin.
2. F. Torrens, G. Castellano (in press) Classification of complex molecules, In: *Foundation on Computational Intelligence*, A.-E. Hassanien, A. Abraham (Eds.), Springer, Berlin.

A COMBINATORICS-BASED DATA-MINING APPROACH TO TIME-SERIES MICROARRAY ALIGNMENT

Turenne N.
INRA, Unité Mathématique Informatique et Génome UR1077, F-78350 Jouy-en-josas, France
e-mail: turenne@jouy.inra.fr

Motivation and Aim: We aim at understanding bovine embryo development and implied proliferation genes. Our approach is based on comparison of two data sets of different species, bovine and human not having the same biological clocks, i.e. protein flows not produced at a same speed for each species. For microarray data the multidimensional property is not handled by 1-D alignment methods. A classical way to process alignment with microarray data concerns univariate time-series curves warping. But it can not be adapted since curves can be deformed if compressed in short-time duration. Hence combinations (unordered partitions) of ordered genes is huge. Our hypothesis aims at aligning some parts of microarrays restricting the hypothesis space of all alignments. We use a clustering approach for each chip and try to merge some clusters by consensus to focus individually on a target gene using symbolic time properties of simultaneity and precedence as described in [1].

Methods and Algorithms: Our method is currently developed under R tool using "clue" library for clustering consensus [2]. First stage relies on gene clustering for each microarray. This part has be done using a k-means method. We obtain resembling gene clusters according their expression profile. At the second stage we try to merge clusters from both microarray into cluster ensembles (CE). A consensus distance is used for merging ("DWH" or fuzzy clustering option). A third stage is dedicated to select, by the user, a given target gene and retrieval of its cluster (CT) it belongs to. At fourth stage we compute a symbolic time correlation matrix of average expression values for each cluster at each time-point and assessing relative occurrence between clusters and CT (B for a cluster occurring before CT, A for after and D for during; a matrix value should be "AD" or "B"). The combinatorial space is expressed in CE and time matrix computation. At firth stage for any gene we can extract the CE in which it belongs and uses it to extract a submatrix of the time matrix. Finally a Jaccard index compares the submatrices obtained for the target gene and any gene to decide if their time profile is similar.

Results: Our methodology describes an original combinatorial approach based on consensus clustering and a symbolic time correlation matrix. We build temporal profile for a given target gene across two time-series microarrays for mining similarity with another gene even if they do not share common microarray occurrence. On a 2 Ghz clock processor the method takes only 11.79 seconds for two sets each one composed of 600 genes and 15 clusters.

Conclusion: Comparing two time series microarray from two different species addresses a question for correlation between time-points and gene comparison across the microarrays since time is not linear between species. All combinations of genes occurrence (ordered partitions) between time-points are possible and huge. We propose a fast methodology based on merged clusters and use of a symbolic time correlation matrix to compute a time profile over two microarrays.

References:

1. N. Turenne and S.R. Schwer (2008) ″Temporal Representation of Gene Networks″, Journal of Data Mining and Bioinformatics (JDMB), Vol.2(1).
2. K. Hornik (2005) "A CLUE for CLUster Ensembles", Journal of Statistical Software, 14(12).

THE RELATIONS BETWEEN CYCLIN/CDK AND HOUSEKEEPING APPARATUS ACTIVITY IN THE CELL CYCLE CONTROLLING: MATHEMATICAL MODELING

Turnaev I.I.*, Gunbin K.V., Likhoshvai V.A.
Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
e-mail: turn@bionet.nsc.ru
* Corresponding author

Motivation and Aim: It is well known that in the cell cycle there were periodicities in: (1) cyclin/CDK protein complex activity [1]; (2) concentration of the key housekeeping metabolites, for example, ATP [2-4]. Therefore, the relations between cyclin/CDK and ATP activities in cell cycle controlling can provoke biological interest.

Methods and Algorithms: Protein sequence homology analysis was performed using PSI-BLAST for the NCBI Genbank nr data (release 161.0), NCBI CDD (version 2.13) and Pfam (release 23.0) data. MGS-modeller was used for mathematical modeling.

Results and Conclusion: It was previously reported that the cell division controlling mechanisms in the pro- and eukaryotes is convergent [5]. Our analysis of protein sequence homology confirmed these previous inferences (Table 1). At the same time our analysis provided that housekeeping proteins of the pro- and eukaryotes share common ancestry (Table 1).

Table 1. The protein sequence homology		Homology % Eukaryote/Prokaryote/ Archae		
Gene group	Representative proteins	Eukaryota	Prokaryota	Archae
cell division controlling mechanisms	CDC13, S.pombe	self homol.	no homol.	27%
	CDC2, S.pombe	self homol.	31%	30%
	CDC25, S.pombe	self homol.	no homol.	no homol.
	CtrA, C. crescentus	30%	self homol.	no homol.
housekeeping proteins	RNA Polimerase B1, S.pombe	self homol.	48%	48%
	DNA Polimerase 2 , E.coli	77%	self homol.	27%
	DNA Polimerase 3 Alpha, E.coli	86%	self homol.	no homol.
	DNA Polimerase Delta, S.pombe	self homol.	71%	31%

Two hypotheses for cell cycle operation were suggested: (1) the periodical activities of housekeeping apparatus is controlled by Cyclin/CDK complexes; (2) the housekeeping apparatus possesses an own oscillator in which ATP plays the key role and this oscillator controls cyclin/CDK protein complex periodical activity. For verification of these hypotheses we are developing the mathematical model, which takes into account Tyson and co-authors data [1] on the yeast cell cycle regulation and regulatory effects of the ATP.

This research may be important for: (1) identification of the relations between the cyclin/CDK activities and the housekeeping apparatus in cell cycle controlling; (2) revealing the evolutionary aspects of these relations.

References:

1. B. Novak, J.J. Tyson (2003) Modelling the controls of the eukaryotic cell cycle, *Biochem. Society Transact.*, 21: 1526-1529.
2. L. Huzik, D.J. Clark (1971) Nucleoside triphosphate pools in synchronous cultures of Escherichia Coli, *J. of Bacteriology* 108, 74–81.
3. G. Orfanoudakis et. al. (1987) Cell Cycle Variations of Dinucleoside Polyphosphates in Synchronized Cultures of Mammalian Cells, *Mol. and Cel. Biol.,* 7, 2444–2450.
4. A.D. Satroutdinov et. al. (1992) Oscillatory metabolism of Saccharomyces cerevisiae in continuous culture, *FEMS Microbiol. Lett.,* 77, 261–267.
5. P. Brazhnik, J.J. Tyson (2006) Cell cycle control in bacteria and yeast: a case of convergent evolution?, *Cell Cycle.*, 5, 522-529.

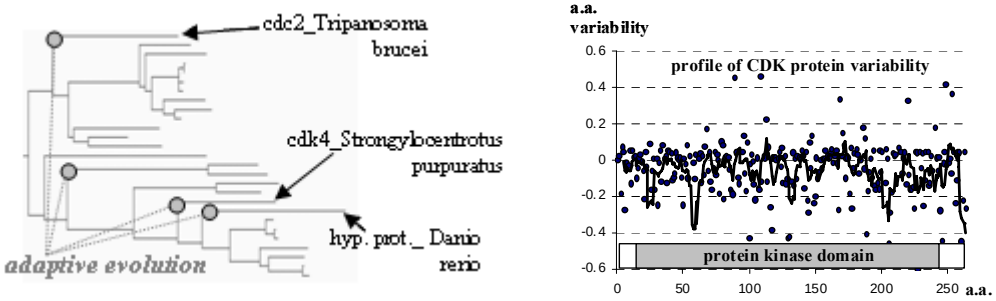
MOLECULAR EVOLUTION OF THE KEY REGULATORY GENES IN THE EUKARYOTIC CELL CYCLE GENE NETWORK

Turnaev I.I.*, Gunbin K.V., Kolchanov N.A.
Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
e-mail: turn@bionet.nsc.ru
* Corresponding author

Motivation and Aim: In the present study, our aim was to investigate molecular evolution of the key regulatory proteins (E2F, pRB, cyclines, CDK, CDI) of the cell cycle in connection with formation of the gene network controlling cell cycle of the multicellular eukaryotes.

Methods and Algorithms: By MAFFT 6.240 [1] we aligned multiply these proteins from GenBank and Ensemble. Phylogenetic trees and ancestor proteins were reconstructed by the maximum likelihood method implemented in PhyML-aLRT 1.1 [2] and FastML 2.02 [3] correspondingly. Branches of the phylogenic trees with positive selection were searched by the computer test based on modeling of protein evolution described in [4] and by comparison of the reconstructed proteins. Detection of protein regions evolving under positive selection was made using Rate4Site 2.01 [5].

Results and Conclusion: We observed that gene duplications are the main mechanism of the cell cycle gene network complexity extension and that positive selection events relate with duplication and major taxon divergence. We although found that regions of proteins evolving under positive selection responsible for regulatory of cell cycle. The fragment of the CDK family tree contained branches with positive selection exemplified at the Fig. 1a; Fig. 1b - protein regions under positive selection. That investigation allows us to propose that cell cycle gene network of the multicellular eukaryotes is formed by consecutive duplication events of regulatory genes following by their positive selection.



Availability: The detailed results are available from the authors upon request.

References:

1. K. Katoh et al. (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.*, **33**: 511-518.
2. M. Anisimova and O. Gascuel (2006) Approximate likelihood ratio test for branchs: A fast, accurate and powerful alternative. *Syst. Biol.*, **55**: 539-552.
3. T. Pupko et al. (2002) A branch-and-bound algorithm for the inference of ancestral amino-acid sequences when the replacement rate varies among sites: Application to the evolution of five gene families. Bioinformatics **18**: 1116-1123.
4. K.V. Gunbin, et al. (2007) The evolution of the Hh-signaling pathway genes: a computer-assisted study, *In Silico Biol.*, **7**: 333-354.
5. I. Mayrose et al. (2004) Comparison of site-specific rate-inference methods: Bayesian methods are superior. *Mol. Biol. Evol.*, **21**: 1781-1791.

COMPUTER-ASSISTED ANALYSIS OF SKIN THERMAL HETEROGENEITY IN HUMANS

Vainer B.G.*, Moskalev A.S., Sapetina A.F.
Rzhanov Institute of Semiconductor Physics, SB RAS, Novosibirsk, Russia
e-mail: bgv@isp.nsc.ru
* Corresponding author

Motivation and Aim: It is recognized and substantiated recently that the extent of thermal heterogeneity of human skin can serve as an independent indicator and a measure of homeostasis shift [1]. 2D thermal pattern (thermogram) of skin is usually visualized and digitized by means of IR thermography. Heterogeneity is presented in thermograms as intensive spottiness. Thermal heterogeneity as a quantitative characteristic is a new feature of living object because firstly it became available since the high-sensitive IR cameras became available, and secondly, irrespective of the method of registration, it is a hidden characteristic of the organism because it declares itself, in the majority of cases, when the organism is loaded. Experimental and computational investigations of inhomogeneous human skin thermal pattern lability is an interesting scientific problem because this novel general feature of the organism discovered exclusively through the use of high-tech equipment can be conditioned and governed partly by human genom. The aims of the present paper are both to throw light on the question concerning a quantitative measure of the extent of 2D thermal pattern heterogeneity, and to develop the computer-assisted automatic method meant for quantitative analysis of inhomogeneous thermograms. The latter allows invoking the computer to facilitate discernment of highly heterogeneous images which are incapable to be undergone a direct analysis.

Methods and Algorithms: Analytic description of the hot areas is made using the features of a normal distribution function. A new numerical technique based on the "water shading" conception is developed and applied to simulated and measured thermograms.

Results: Two examples of diverse degrees of skin temperature heterogeneity shown up under physical load, and the results of corresponding automatic retrieval of warm areas (spots) are presented in fig. 1. It is evaluated that, at present, the method is accurate to about 8 %. The process of the "diffluence" of the spots in time is also investigated.

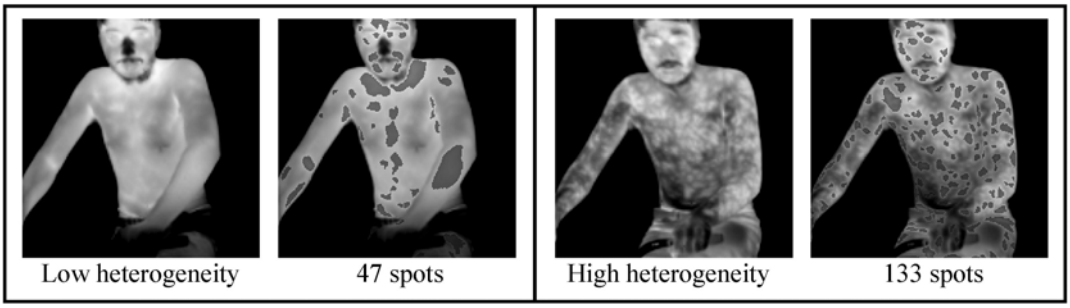


Fig. 1. Low- and highly heterogeneous thermograms, and the retrieved hot spots.

Conclusion: The developed method gives a tool for adequate quantitative analysis of heterogeneous thermal patterns as well as the patterns of different biological and other objects characterized by surface inhomogeneity. It offers a clearier view of how the lability of the patterns and the human's condition or its genom can be interrelated.

References:

1. B.G.Vainer (2005) FPA-based infrared thermography as applied to the study of cutaneous perspiration and stimulated vascular response in humans, *Physics in Medicine and Biology*, **50**: R63–R94.

GENOME-WIDE ASSESSMENT OF THE CODON USAGE CONSERVATION

***Vinogradov D.V.*[1*], *Mironov A.A.*[1, 2]**
[1] Institute for Information Transmission Problems, Moscow, Russia
[2] Moscow State University, Moscow, Russia.
e-mail: dimavin@bioinf.fbb.msu.ru
* Corresponding author

Motivation and Aim: Unequal usage of synonymous codons has been proved to impact the protein folding or the gene product expression rate significantly in some cases (e.g. [1], [2]). The common belief is that such cases are widespread, though no genome-wide analysis of codon usage conservation was published. Under the basic assumptions, functionally important properties should be conserved in orthologous groups. The goal of this research is to check whether the codon usage is conserved during the evolutionary process.

Materials and Methods: A set of manually curated multiple protein alignments was taken from the PFAM database (PFAM-A), then corresponding DNA alignments were built using the UniprotDB and EMBLCDS databases. Relative frequencies of codons for all species were calculated using CUTG database, and each line of every alignment was converted to a series of these frequencies. The averaged correlation coefficient for these series was used as a conservation measure. The probability to observe such correlation coefficient was calculated by a Monte-Carlo simulation of a codon selection among possible synonyms for a given amino acid.

Results: We found that in our data set the average CC of codon frequencies series is about 40%. This value is comparable to that of the Monte-Carlo simulation, which means the codon usage distribution is mainly random. This fact allows us to build a probabilistic model for the synonymous codons distribution and to define more clearly the cases when codon usage in a gene region cannot be explained by purely stochastic reasons.

References:

2. C. Kimchi-Safraty et al. (2007) A silent polymorphism in the MDR1 gene changes substrate specificity, *Science,* **315**: 525-528
3. Y.M. Zalucki, M.P. Jennings (2007) Experimental confirmation of a key role for non-optimal codons in protein export, *BBRC*, **355**: 143-148

MODIFIED DNA COMPLEXES AS BUILDING BLOCKS FOR NANOBIOENGINEERING

Vinogradova O.A.[1*], Lomzov A.A.[1, 2], Rodyakina E.E.[3], Latyshev A.V.[3], Klinov D.V.[4], Pyshnyi D.V.[1, 2]

[1] Institute of Chemical Biology and Fundamental Medicine, SB RAS, Novosibirsk, Russia
[2] Novosibirsk State University, Novosibirsk, Russia
[3] Institute of Semiconductor Physics, SB RAS, Novosibirsk, Russia
[4] Institute of Bioorganic Chemistry, Moscow, Russia
e-mail: viola@niboch.nsc.ru
[*] Corresponding author

Motivation and Aim: Over the last few years self-assembling DNA-based systems are considered as promising building blocks for nanobioengineering, namely for the design of DNA nanoobjects of varied forms, molecular machines and nanodevices. One of the problems in the design of DNA-based architectonics is the high rigidity of the double stranded DNA structure. The aim of this work was to study a possibility of the use of chemically modified DNA fragments as building blocks for nanodesigning atypical dsDNA structures with definite geometrical characteristics.

Methods and Algorithms: Oligonucleotides were synthesized on ASM-700 DNA synthesizer. The structure of supramolecular DNA complexes was investigated by methods of gel-retardation and atomic-force microscopy (AFM).

Results: The influence of the internal loop of different nature and length on DNA complexes was investigated in respect of their conformational features and stability. The electrophoretic mobility of the native DNA duplex and complexes containing various extrahelical bulges in the middle of duplexes was studied. These bulges were formed by introduction of nucleotide or non-nucleotide inserts based on phosphodiester of diethylene glycol or decandiol. The magnitudes of the DNA helix bending at the site of perturbation were calculated. The influence of both the nature (nucleotide or non-nucleotide) and the length of the inserts on the bending angle value was determined. Using modified DNA blocks and concatemeric structures, non-typical dsDNA constructions were designed. It was revealed that the relative orientation of the bended units (torsion angle) within the concatameric structures affect the features of self-assembling DNA nanoobjects. The structure of designed concatamers was confirmed by atomic-force microscopy (AFM).

Conclusion: The use of modified DNA duplexes opens the opportunity to compact DNA-based nanoconstructs with the pre-determined geometry and the sufficient stability.

This work was supported by integration grants of SB RAS (55, 73), by Program MCB of RAS, by CRDF (RUX0-008-NO-06, Y2-B-08-03), RFBR (06-04-49263) and by Program UMNIK.

ANALYSIS OF THE DEGENERATE MOTIFS IN 5'-REGULATORY REGIONS OF PROCARYOTES

Vishnevsky O.V. [1, 2]
[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090 Russia
[2] Novosibirsk State University, ul. Pirogova 2, Novosibirsk, 630090, Russia
e-mail: oleg@bionet.nsc.ru

Motivation and Aim: In spite of the success of experimental biology in the analysis of prokaryotic gene regulatory regions (GRR) and in the new transcription factor binding sites (TFBS) revealing, a lot of important details of GRR organization are still unclear.

The aim of our research is to develop a new approach for recognition and classification of the regulatory signals in the prokaryotic genomes.

Methods and Algorithms: The sets of *B.subtilis*, *E.coli*, *H.pylori*, *M.gallisepticum*, *M.genitalium* and *M.pneumoniae* GRRs in the [-100; +25] region relative to the translation start are created. We developed a new approach for detection of degenerate (written in extended 15 single letter-based IUPAC code) region-specific oligonucleotide motifs in full-genome sets of prokaryotic GGRs. The degenerate oligonucleotide motif obtained using this approach is considered significant if its occurrence in the GRR sample is more than 10% and its binomial occurrence probability is lower than 10^{-8}.

Results: A few hundreds (N_{motifs}) of significant degenerate motifs for every set are obtained using our system. It was found, that N_{motifs} does not correlate with the number of the sequences in the sets. Then we classify the motifs by using database of *E.coli* TFBS. A different number (N_{TFBS}) of the motifs, significantly ($p<10^{-4}$) presented in the sets of *E.coli* TFBS are categorized as potential TFBS for every GRR set. As expected, the biggest N_{TFBS} was obtained for motifs detected in *E.coli* GRRs. We suppose the TFBS specific motifs, obtained in other species could bind to transcription factors (TF) homologous to the *E.coli* TFs. All other motifs could correspond to the species-specific TFBF, those absent in *E.coli* database or to some structural features of prokaryotic GRRs, like short polyA-polyT runs similar to those found in the GRRs are known to induce DNA curvature or to be "easily melting" sites. We compare behavior of regulatory motifs in GRR of evolutionary close and far prokaryotic species using the approach we offered [1] to estimate the presence and distribution of oligonucleotide motifs W_{voc} in GRR. W_{voc} is calculated for GRRs of all species considered, basing on the motifs specific for *M.genitalium*. At the same time, an average similarity (H_{align}) between GRRs of the same species and *M.genitalium* GRRs is estimated using pairwise alignment. It was shown that the W_{voc} level is much higher for related to *M.genitalium* species (like *M.gallisepticum* and *M.pneumoniae*) than for evolutionary far species (*E.coli*) at the similar H_{align} level.

Conclusion: We revealed potential transcription factor binding sites in regulatory regions of 6 species of prokaryotes that can be an aim of the experimental analysis.

It was shown that regulatory regions of evolutionary related species are more similar in terms of the functional signals, than in terms of an average homology.

References:

1. O.V. Vishnevsky, N.A.Kolchanov (2005) ARGO: a web system for the detection of degenerate motifs and large-scale recognition of eukaryotic promoters, *Nucleic Acids Research*, **33**, 417-422, Web Server issue

GENETIC DIVERSITY INVESTIGATION AND PASPORTIZATION OF TRIBE VICIEAE (ADANS.) BRONN REPRESENTATIVES FROM VIR COLLECTION BY MEANS OF RAPD-ANALYSIS

Vishnyakova M.A.*, Burlyaeva M.O., Alpatieva N.V., Chesnokov Yu.V.
Vavilov Institute of Plant Industry, Saint-Petersburg, Russia
e-mail: m.vishnyakova@vir.nw.ru
* Corresponding author

Motivation and Aim: Molecular markers are a very suitable tool for gene bank management for different purposes: identification of genotypes, the resolution of systematic and phylogenetic problems, searching duplicates and pasportization of the accessions. The establishment technically optimized collections well documented DNA samples also becomes a question of vital importance in world genebanks. Investigation and pasportization of representatives tribe *Vicieae* with the aim of RAPD-analysis and collecting DNA samples is the aim of the research. The tribe is known as the most perfect in the family *Fabaceae* Lindley, but having a lot of taxonomic problems and obscure phylogeny. Many representatives of the tribe, preserved in VIR collection, have a great economic value.

Methods and Algorithms: The sample of 250 representatives of 51 species from the tribe *Vicieae* belonging to 10 genera (*Orobus* L., *Bona* Medik., *Faba* Mill., *Vicia* L., *Ervum* L., *Lens* Mill., *Ervilia* (L.) Link., *Pisum* L.*, Lathyrus* L.*, Clymenum* Mill.) had been created. The accessions selected had the known morphological characteristics, represented the most species types and reflected natural area. In total 500 DNA samples had been got and subjected to RAPD analysis. 22 primers were applied from which 9 giving the highest polymorphism were chosen for further work: OPA10, OPH2, OPH3, OPH6, OPH9, OPK4, OPK8, OPK9, OPK10. The special attention had been paid to the genus *Lathyrus* from which 37 species had been investigated.

Results: With the selected primers sufficient polymorphism had been detected allowing identification of each 250 accessions. Intra- and intergeneric level of polymorphism had been determined. Significant diversity of wild as well as cultivated representatives of tribe *Vicieae* had been shown. Specific fragments for genera and species studied have been found. For the establishment of intergeneric relationships in the tribe 94 DNA samples of 38 species had been selected which had been amplified with 4 RAPD primers (OPA10, OPK4, OPK9, OPH3). 226 polymorphic DNA fragments have been analysed and dendrogram of genetic similarity elaborated.

On the bases of data obtained some disputative questions of phylogeny and systematic of the tribe *Vicieae* and taxonomy of the genus *Lathyrus* became more clear.

As the result of the research the collection of 500 DNA samples of accessions with different origin and level of domestication is created. The passport DB of these samples had been elaborated.

Conclusion and Availability: RAPD-analysis is a rapid and reliable method for genera and species discrimination and establishment of their relationships. Passport DB of accessions of tribe *Vicieae* created on the basis of RAPD-analysis will be submitted on-line on VIR web-site. 500 DNA samples became a contribution to VIR DNA bank for short-term or long-term storage.

The research is supported by grant of RFBR 06-04-48869-a

THE CATARACTOGENIC EFFECT OF MUTATIONS IN THE CRYSTALLINES MAY BE COMPENSATED BY SUBSTITUTES IN A SYMMETRIC DOMAIN

Vlasov P.K.
Engelhardt Institute of Molecular Biology RAS, Moscow, Russia
e-mail: vlasov@imb.ac.ru

Motivation and Aim: Crystallins are the major structural proteins in the mammalian eye lens and consist of different families with the γ-crystallin composing up to 40% of the soluble proteins expressed in the lens [1, 2]. A new mutation (P23S) was identified in the gene of the γD-crystallin that is associated with a polymorphic congenital cataract in a human population [3]. Nevertheless, Ser23 corresponds to the normal state of gene in a lower primate and all the surveyed nonprimate mammals. In our work we analysed possible compensatory substitutes for the cataractogenic effect of Ser23.

Results: We undertook a correlation analysis in search of compensatory substitutions for Ser23 in the entire protein, using a multiple alignment of all sequences of the γ-crystallin mammalian proteins. Site 23 interacts with position 49, and, since many compensatory substitutions in interacting sites have been described, we surveyed the amino acid at site 49 in mammalian γD-crystallins. However, we found that site 49 and the neighboring sites are generally conserved throughout evolution and show no evidence of compensatory evolution with site 23. By the way, we found that substitutions at site 23 were always associated with substitutions in the sites 109 and 136. Remarkably, the protein forms two structurally similar domains, and site 109 corresponds to the same position in the second domain as site 23 in the first domain, whereas the corresponding site of the site 49 in the second domain is site 137 – naturally, the neighbor of the site 136.

Conclusion: Thus, it is likely that the P23S substitution in the N-terminal domain is compensated for in another part of the protein in a distal domain. This effect may be caused by a change in the hydrogen-binding characteristics of the protein-water interface. A substitution of a proline, since it is an imino acid that does not have a hydrogen bond–forming NH group, is particularly capable of affecting protein solubility in water.

References:

1. J.Craw (1997) The crystallins: genes, proteins and diseases. *Biol Chem,* **378:** 1331-1348.
2. S.O.Meakin et al. (1987) γ-Crystallins of the human eye lens: expression analysis of five members of of the gene family, *Mol Cell Biol*, **7:** 2671-2679.
3. O.V.Plotnikova, F.A.Kondrashov et al. (2007) Conversion and compensatory evolution of the γ-crystallin genes and identification of a cataractogenic mutation that reverses the sequence of the human CRYGD Gene to an ancestral state, *Am J Hum Genet*, **81(1):** 32–43.

POLYMORPHISM OF LIPOPROTEIN LIPASE GENE IN WEST SIBERIA CAUCASIAN POPULATION AND ITS ASSOCIATION WITH PLASMA LIPID LEVELS

Voevoda M.I., Shakhtshneider E.V.*, Kulikov I.V., Maksimov V.N., Romashchenko A.G., Nikitin Yu.P.
Institute of Internal Medicine, Novosibirsk, Russia
e-mail: sch1@rbcmail.ru
* Corresponding author

Aim: We investigated polymorphism of lipoprotein lipase gene (LPL) and its influence on plasma lipids levels in Caucasian population of West Siberia.

Methods: The patients included in the analyses were selected based on total cholesterol (TC) level from population sample surveyed in frame of HAPIEE project (~9000 participants, aged 45-69, men 50%). Totally 100 patients with total cholesterol level (TC) >300mg/dl, 100 patients with TC<200mg/dl and 100 patients with TC corresponding to population mean - 233.6±47,7mg/dl were included in the analyses. The plasma lipids levels were determined by standard enzymatic assays. All patients were of Caucasian origin. The 22125T/G (Hind III) polymorphism of LPL gene was analyzed by standard method.

Results: The frequencies of H- and H+ alleles in patients with highest TC level were 34.8% and 65.2% respectively. The frequencies of H- and H+ alleles in patients with lowest TC were 29.6% and 70.4%. The frequencies of H- and H+ alleles in patients with intermediate TC total cholesterol level were 21.7% and 78.3%. The frequencies of genotypes H-H-, H+H- and H+H+ were 15.9%, 37.8% and 46.3% in patients with highest TC level. The frequencies of genotypes H-H-, H+H- and H+H+ were 9.1%, 40.9%, 50.0% in patients with lowest TC level. The frequencies of genotypes H-H-, H+H- and H+H+ were 3.0%, 37.4%, 59.6% in patients with intermediate TC level. The differences of TC, triglycerides and high density lipoproteins cholesterol levels between genotypes in patients with highest total cholesterol level are not significant. The differences of TC level in case of genotypes H-H- are significant in group of patients with lowest TC level and in group patients with intermediate TC level. TC level in case of genotypes H-H- is highest. Mean total serum cholesterol levels in case of genotypes H-H-, H+H- and H+H+ were 308.9±23.1mg/dl, 249.0±11.0mg/dl, 250.2±9.4mg/dl (pANOVA=0.05). The differences of triglycerides and high density lipoproteins cholesterol levels between genotypes are not significant.

Conclusions: The Caucasian population of West Siberia is not significantly differs from populations of Europe and North America by frequencies of alleles and genotypes. The genotype H-H- has been associated with higher total serum cholesterol level in comparison with genotypes H+H- and H+H+ in group patients with lowest total cholesterol level and in group patients with populations total cholesterol level.

This work was supported by Grant of Federal Agency of Science of Russia №02.442.11.7515, 2006.

FINE STRUCTURE OF MAMMALIAN TRANSLATION INITIATION SIGNAL

***Volkova O.A.*[1*]*, Kochetov A.V.*[1, 2]
[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
[2] Novosibirsk State University, Novosibirsk, Russia
* e-mail: ov@bionet.nsc.ru

Motivation and Aim: AUG initiation codon recognition efficiency depends on its context. Guanine in +4 position was proposed to increase translation initiation efficiency especially when pyrimidine occupied position –3. 3'-end segment of AUG context is a part of two functional regions: both translation initiation signal and protein coding sequence. It was shown that the content of amino acids in N-terminal CDS positions is considerably biased. The most frequent N-terminal amino acids are alanine and serine. It was assumed that these amino acids could increase protein stability and facilitate posttranslational protein modification. However, these amino acids could also facilitate translation as a part of translation initiation signal. In this work we analyzed the interdependency between start codon context and N-end amino acids.

Materials and Methods: The mRNAs of *Mus Musculus,* and *Homo sapiens* genes were extracted from the EMBL database, homologous sequences were removed with the aid of the CleanUp http://www.ba.itb.cnr.it/BIG/CleanUP/. The resulting sets included 16788 mRNAs of *Mus musculus* and 24154 of *Homo sapiens*. Statistical analysis was made with Statistica program (StatSoft^{TM}, Inc.). Estimation of statistical significance of deviation between sets accomplished according to *t*-test. Sets were divided into subsets depends on nucleotide in the-3 position (A^{-3}, G^{-3} or Py^{-3}).

Results: We analyzed subsets of human and mouse mRNAs containing in position –3 A, G or Py on the frequencies of nucleotides in CDS positions from +4 to +6 and amino acids in protein position number 2. It was found the mRNAs with A^{-3} and G^{-3} differ significantly: G^{-3} positively correlated with the presence of G^{+4} and Ala^{2}, whereas A^{-3} negatively correlated with G^{+4} and positively correlated with Ser^{2}.

Conclusion: We found that G^{+4} could be an important part of GnnAUG translation initiation signal. We also assumed that Ala and Ser in 2nd protein position could facilitate translation initiation at signals of different types (AnnAUG and GnnAUG, respectively).

Acknowledgments: This work was supported by RFBR (08-04-00525) and the RAS program “Dynamics of Gene Pools”. We also thank SD RAS (grant No. 5.3) for partial support.

PREDICTION OF SPATIAL STRUCTURE OF TRANSMEMBRANE HELICAL DIMERS USING MOLECULAR MODELING TECHNIQUES

Volynsky P.E.*, Nolde D.E., Efremov R.G.
Institute of Bioorganic Chemistry RAS, Moscow, Russia
e-mail: pashuk@nmr.ru
* Corresponding author

Motivation and Aim: Receptor tyrosine kinases (RTK) represent an important class of membrane proteins. They participate in such key processes as cell differentiation, growth of tissues and rearrangements in induced by external signals (ligand binding). RTKs consist of three parts – extracellular ligand-binding domain, cytoplasmic kinase domain and helical transmembrane linker fragment (TMS). The TMS takes part in RTK dimerization along with the extracellular domain. Moreover, it was found that some diseases are caused by mutations in TM fragments of RTK which lead to their permanent activity, metabolic disorder and cell death. Information about structure of TMS dimers helps in understanding of the principles of RTK activity. Moreover, it may be used in construction of selective medicine to diseases, caused by dysfunction in RTK dimerization. Experimental solution to these problems is very difficult because of a complex nature of RTK environment. All these facts promote development of approaches for fast computational prediction of the structure of dimeric transmembrane helices.

Methods and Algorithms: Prediction of a dimeric structure is based on a combination of computational techniques with different degree of approximation. The simplest approach is analysis of hydrophobic properties of TMS surface. Usually, this method produces several crude models of dimeric structure. More equilibrated models can be derived using Monte Carlo conformational search in implicit membrane. In addition to the structure of dimer, this approach also permits identification of the geometry of the dimer in membrane. And, finally, these models may be refined via molecular dynamics in explicit hydrated bilayer. The last two approaches can be easily adapted to parallel calculations, thus decreasing time cost of the computation. Combined application of these different approaches to analysis of TMS dimers provides relatively good models of dimeric structures.

Results: The aforementioned methods were applied for investigation of the spatial structure of some test dimers of TMS with experimentally determined structure (glycophorin A, bnip3). Comparison of the simulation results with experimental data permits delineation of advantages and disadvantages of different approaches. It should be noted that in each case the computational models were in a good agreement with the experimental ones. Then, simulations of dimers of TMS of biologically important RTK with unknown structure were carried out.

Conclusion: Good agreement of simulation results with the experimental information validates the proposed approach to prediction of spatial structure TMS dimers. This approach can be easily applied for structure prediction of any dimer (without difficulties connected with its experimental investigation). Finally, at least partially, these approaches may be elaborated for solution of more difficult problems, for instance, to development of the aforementioned medicines or prediction of more complex oligomers (bundle of helices, etc.)

A NOVEL EXHAUSTIVE DOCKING METHOD COMBINING CAVE & GROOVE SEARCH WITH GLOBAL MOLECULAR DYNAMICS OPTIMIZATION

Vorobjev Y.N.
Institute of Chemical Biology and Fundamental Medicine, SB RAS, Ac.Lavrentieva Ave.8, Novosibirsk 630090, Russia
e-mail: ynvorob@niboch.nsc.ru

Motivation and Aim: An effective search for a new targets and drugs needs a reliable computational tools for exhaustive docking of ligands on protein molecule. Docking of a flexible ligand on protein is a challenging long standing problem of computational biology. The available docking packages DOCK4.0, FlexX1.8, GOLD1.2, AutoDock3.05 found the native binding mode within RMSD 3.0 Å with success rate (SR) in the range of 40-65 % [1,2]. A nowel effective hierarchical algorithm of exhaustive docking of a flexible ligand on a rigid/flexible protein with reliable scoring method has to be developed and implemented.

Methods and Algorithms: The novel docking method is based on exhaustive analysis of probe accessible surface of biopolymer, while the most of docking packages use a search over 3D-grid around biopolymer [1]. An effective search for caves and grooves over protein molecule allow us to calculate a sparse surface grid of binding site candidates. The found site are roughly scored by number of prlotein-ligand contacts. A final refinement and scoring of binding sites are done for a subset of binding site candidates by a global optimization method. The optimization is based on multiple start for different ligand orientations and coupled force field variation and temperature annealing by the method of molecular dynamics. The main stages of docking algorithm are: 1) calculation of a probe accessible surface of protein; 2) analysis of the PAS and calculation of binding site surface grid as a virtual positions which are able to accommodate a typical chemical groups; 3) estimation of a contact surface complementary score; 4) calculation of simplified ligand image; 5) global (rotational/translational) optimization of ligand image on the docking surface grid; 6) global optimization of the ligand via full atom molecular dynamic simulating annealing with variable force field with flexible ligand and rigid/flexible protein.The final binding score of different binding sites are ranged according to average energy of protein-ligand interactions with modified force field calculated for flexible ligand by method of molecular dynamics at temperature of 50 K.

Results: Docking method is tested on a variety of protein-ligand complexes (> 30) with ligands size ranged from small rigid molecule, like benzamidine to a large flexible ligands like, agrotroban and VAC inhibitor of HIV1 protease, ligands of peptide nature, i.e. 5-, 8-residue-peptides and dinucleotides. The new docking method identifies the native binding mode as the mode with highest binding score, estimated at the final stage of molecular dynamic optimization protocol, with success rate about 90%.. The total CPU demand for exhaustive docking search is about 10-12 hours per one Pentium IV (~ 2 GHz) processor, per one complex. The presented docking algorithm is suitable for coarse grain parallelization.

Conclusion: The presented docking method show a high rate of successful prediction of the native binding mode with rmsd < 3 Å for variety of protein-ligand complexes. The testing is performed for rigid protein in experimentally available protein-ligand complexes. The method is implemented as a modul of the BISON program package for biomolecular simulations .The final stage of ligand optimization can be done for a flexible protein&ligand in the complex.

The work was supported by the Russian Foundation for Basic Research (projects no. 05-04-48322 a and 08-04-00327a).

References:

1. Hung-Ming Chen et al, (2007) J Comp.Chem. **28**:612-623.
2. Ruvinsky A.M. (2007) J Comp Chem. **28**:1364.

MODELING OF ATOMIC STRUCTURE OF MULTIMOLECULAR COMPLEXES INTEGRATING CALCULATIONS WITH CHEMICAL CROSSLINKING DATA

Vorobjev Y.N.*, Kiselev L.L.
Institute of Chemical Biology and Fundamental Medicine, SB RAS, Ac.Lavrentieva Ave.8, Novosibirsk 630090, Russia
e-mail: ynvorob@niboch.nsc.ru

Motivation and Aim: 3D atomic structures of eukaryotic ribosomal complexes are unknown. A modern understanding of mechanism of translation assumes its understanding in terms of 3D atomic structure of complex. eRF1•mRNa• tRNA containing human class-1 polypeptide release factor eRF1 at the A-site of human 80S ribosome, mRNA and P-site tRNA. A computational modeling becomes the only tool to obtain 3D models of such complex.

Methods and Algorithms: A multi step modeling of a structure of triple complex based on structures of individual molecules of the complex is developed. Method consist of steps: 1) investigation of possible conformational rearrangements of individual molecules by analysis of essential conformational movements via method of molecular dynamics, 2) building of a preliminary models of complex using a variant of method of comparative modeling and docking, 3) refinement of models via modeling of biochemical data. The X-ray structure of the T. thermophilus 70S ribosome containing $tRNA^{Phe}$ in the P and A sites and the crystal and high-resolution NMR structure of the human eRF1 are used as initial data. Atomic model has been constructed via protocol: (1) determination of an essential conformational movements of domains of isolated eRF1,) a deformation of eRF1 structure with the aim to fit the mutual arrangement of the N and M domains to the shape of template A-site tRNA; (3) substitution of the A-site tRNA in the T. thermophilus 70S ribosome with eRF1 in a tRNA-like conformation; (4) docking of eRF1 in the tRNA-like conformation to mRNA nucleotides from +3 to +9 relative to the first nucleotide of the UUC codon in the P site of the 80S ribosome, and (5) optimization of position and structure of eRF1, taking into account all interactions with mRNA and the neighbor P-site tRNA and constrains ensuring a proper spatial positioning of the GGQ motif relative to the 3'-CCA end of the P-site tRNA. The method yielded two structural models of the complex . To choose between the models we used the data on chemical crosslinking between mRNA and eRF1 in the A site obtained with 12 photoactivatable mRNA analogs carrying a reactive group at nucleotides in positions from +4 to +9. A probability of chemical crosslinking of the reactive group to eRF1 residues in models 1 and 2 of the triple complex have been estimated via modeling. The modeling has been done via our program complex BISON .

Results: The distribution of short contacts (<7 A) between the azido group of modified nucleotides and atoms of the N- and C-domain residues of eRF1 agrees well with the experimental results of chemical crosslinking for all 12 structures of mRNA analogs only for model 1, which can be identified as the final state of ribosomal termination complex.

Conclusion: A novel methods for building and validation of 3D structure of macromolecular complexes has been developed. A model of unknown structure of the final state of eukaryotic ribosomal triple complex eRF1•mRNa•tRNAPhe is obtained and validated by comparison the calculated and experimental chemical crosslinking data. The developed methods extend a set of tools for structural computational biology. Obtained results will facilitate better understanding of functions of release factor eRF1.

The work was supported by the Russian Foundation for Basic Research (projects no. 05-04-48322 and 06-04-48037).

ANALYSIS OF FACTORS AFFECTING THE ACCURACY PREDICTIONS FOR PROTEIN-PROTEIN INTERACTIONS BASED ON THE MIRROR TREE APPROACH

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

Motivation and aims: A promising approach to computer prediction of protein-protein interactions is the mirror tree approach, which is based on analysis of similarities among the phylogenetic trees of protein families [1]. It is suggested that, for interacting proteins, due to their co-evolution, the similarities of trees for interacting proteins are greater than the non-interacting. A limitation of the method is the great number of false positive predictions because the phylogenetic tree of a protein family as a rule significantly correlates with the phylogenetic tree for organisms [2].

Here we analyzed the various factors, which can affect the accuracy predictions for protein-protein interactions using the mirror tree approach. Under study were the conserved positions of multiple alignments, the hydrophobisity index and other physico-chemical properties of multiple alignments, the additional similarity measures between phylogenetic trees using Kendall and Stuart correlation coefficients and partial correlation coefficients.

Methods and Algorithms: The prediction results were tested on a sample of 178 *E.coli* proteins extracted from the DIP database [3], homologous sequences were taken from the KEGG database [4]. Sequences aligned using the Mafft software [5], the Protdist program from the Phylip package [6] was used for building the matrix of evolutionary distances. The conservation of the positions was analyzed utilizing the CRASP program [7].

Results: This analysis demonstrated that additional consideration of the above factors allows increasing the prediction accuracy of protein-protein interactions. Thus, when the evolutionary features of conserved and variable positions of protein families are taken into account, the proportion of false positives predictions is considerably decreased.

The work was supported by the RAS programs “Biosphere origin and evolution” and “Molecular and cellular biology” and SB RAS integration projects #115 and #49.

References:

2. Valencia A, Pazos F (2002) Computational methods for the prediction of protein interactions, *Curr Opin Struct Biol*., **12**: 368-373.
3. Sato T, Yamanishi Y, Horimoto K, Kanehisa M, Toh H. (2006) Partial correlation coefficient between distance matrices as a new indicator of protein–protein interactions, *Bioinformatics*, **22**: 2488-2492.
4. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D (2004) The Database of Interacting Proteins, *Nucleic Acid Res*, **32**: D449- D451.
5. Kanehisa, M. and Goto, S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes, *Nucleic Acids Res*., **28**: 27-30.
6. Katoh, Misawa, Kuma and Miyata (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform, *Nucleic Acids Res*., **30**: 3059-3066.
7. Felsenstein, J. (1989) PHYLIP-phylogeny interference package, *Cladistics*, **5**: 164-166.
8. Afonnikov DA, Kolchanov NA. (2004) CRASP: a program for analysis of coordinated substitutions in multiple alignments of protein sequences, *Nucleic Acids Res*., **32**: W64-W68

THE PARALLELIZATION OF THE PLATO ALGORITHM FOR ANALYSIS OF THE ANOMALOUSLY EVOLVING GENE REGIONS

***Vyatkin Yu.V.*[1*]*, Gunbin K.V.*[1]*, Snytnikov A.V.*[2]*, Afonnikov D.A.*[1, 3]
[1] Institute of Cytology and Genetics, SB RAS, Lavrentyev aven. 10, Novosibirsk, Russia
[2] Institute of Computational Mathematics and Mathematical Geophysics, SB RAS, Lavrentyev aven. 6, Novosibirsk, Russia
[3] Novosibirsk State University, Pirogova Str. 2, Novosibirsk, Russia
e-mail: vyatkin@bionet.nsc.ru
[*] Corresponding author

Motivation and Aim: An important issue in comparative analysis of genomic sequences from different organisms is detection of genes or their parts that possess specific modes of nucleotide substitutions. Grassly and Holmes [1] have proposed a method for the detection of gene regions evolving anomalously using the maximum likelihood approach implemented in the PLATO program. We performed parallelization of the PLATO algorithm and applied the modified program to analysis of gene sequences of the myostatin family [2].

Methods and Algorithms: The identification of the anomalously evolving regions in the PLATO program is based on the likelihood function estimation for a window scanning along a sequence [1]. The calculation of the matrix of likelihood values for each position and window length is the most time consuming part of the algorithm. In our parallel implementation, the matrix element calculation was equally distributed among the processors as jobs. This distribution is done automatically, depending on how many processors are accessible to the program. Thus, each processor contains a piece of the resulting matrix after finishing its job. The pieces are assembled into the matrix, which is then sent to all the processors so that each contains a full copy of the similarity matrix for the sequences being analyzed by the sliding window. Parallelization is done using the MPI library. There is a complete agreement in output results between the PLATO serial and parallel versions.

Results: We analyzed the evolutionary features of the genes of the myostatin subfamily (*GDF-8*) with parallelized version of the PLATO program. The computation demonstrated that the likelihood values are different from region to another for myostatin. The obtained results are consistent with loads imposed on these domains, what may be due to the positive selective pressure, as previously reported [3].

Calculations using the parallel version of PLATO ran at PC-clusters with different numbers of CPUs. The more processors were employed per task, the less time it took to complete the calculations; for example, 256 processors did the job in 15 seconds (for comparison, one processor did it in 45 minutes). Thus, we acquired almost linear speedup of computations depending of number of processors used.

The work was supported by the RAS programs “Biosphere origin and evolution” and “Molecular and cellular biology”.

References:

1. Grassly N C, Holmes EC (1997) A likelihood method for the detection of selection and recombination using nucleotide sequences, *Mol Biol Evol.* 14: 239-247.
2. Tsuchida K (2004) Activins, myostatin and related TGF-beta family members as novel therapeutic targets for endocrine, metabolic and immune disorders, *Curr Drug Targets Immune Endocr Metabol Disord.*, 4: 157-166.
3. Tellgren A, Berglund A C, Savolainen P, Janis C M, Liberles D A (2004) Myostatin rapid sequence evolution in ruminants predates domestication, *Mol. Phylogenet. Evol.*, 33: 782-790.

GENETIC COLLECTION AND DEVELOPMENT OF NEAR-ISOGENIC LINES IN WHEAT

Watanabe N.
College of Agriculture, Ibaraki University 3-21-1 Chuo, Ami, Inashiki Ibaraki 300-0393 Japan
e-mail: watnb@mx.ibaraki.ac.jp

Since Watanabe (1994), more than forty near-isogenic lines were developed in durum wheat cultivar LD222. The use of genetic collections of tetraploid and hexaploid wheats were considered. The genes to be introduced are located on the specific chromosome and mapped in the linkage maps using the aneuploid stocks of LD222 and Langdon, Landgdon D-genome chromosome substitution lines, and microsatellite markers. We contributed the mapping of the genes for long glumes on chromosomes 7AL and 7BL (Watanabe et al. 1996, 1999, 2002; Watanabe & Imamura 2002), brittle rachis on chromosomes 3AS and 3BS (Watanabe et al 2002, 2006) and liglulesness on chromosome 2BL (Watanabe et al 2004). Two mutations for sphaerococcoid seed (MA16219) and compact spike (MA 17648) isolated from M_3 progeny of a durum wheat cultivar, Altaiskaya Niva. The gene for sphaerococcoid grain, s^{16219}, was allelic to *S2*, which is located on the centromeric region of chromosome 3B in hexaploid wheat. The gene for compact spike, C^{17648} is located on the chromosome 5A. It was observed that C^{17648} was different from the *Q* locus. The near-isogenic lines for sphaerococcoid seed and compact spike were established as ANW 22A and ANW 11D. The near-isogenic lines for GA-sensitive *Rht* genes (*Rht 14*, *Rht 16*, *Rht 18* and *Rht 19*) were developed, although their chromosomal locations have not been determined. The multiple alleles at *Rht-B1* locus were introduced into the genetic background of cv. LD222. *Triticum polonicum* IC 12196 may be considered as new source of *Rht* gene (Watanabe, 2002). The effort to develop near-isogenic lines was extended to introduce taxonomy-related traits such as spelt, squarehead and awn on the glumes. Several near-isogenic lines are available upon request.

References:

1. Kosuge K, Watanabe N, Kuboyama T, Melnik VM, Yanchenko VI, Rosova MA, Goncharov, NP (2007) Euphytica 1**59**: 289-296.
2. Watanabe, N (1994) Euphytica 72:143-147.
3. Watanabe, N, Yotani, Y, Furuta, Y (1996) Euphytica 90: 235-239.
4. Watanabe, N (1999) Euphytica 106: 39-43.
5. Watanabe N, Sekiya T, Sugiyama K, Yamagishi Y, Imamura I (2002) Euphytica 28:129-134.
6. Watanabe N, Sugiyama K, Yamagishi Y, Sakata Y (2002) Hereditas 137: 180-185.
7. Watanabe N, Imamura, I (2002) Euphytica 128: 211-217.
8. Watanabe N, Nakayama A, Ban T (2004) Euphytica 140: 163-170.
9. Watanabe N (2004) Cereal Res. Commun. 32: 429-434.
10. Watanabe N, Fujii Y, Kato N, Ban T, Martinek P (2006) J. Appl. Genet. 47: 93-98.

FEATURE SUBSET SELECTION FOR CANCER CLASSIFICATION USING MAXIMIZED MARGIN OF SUPPORT VECTOR MACHINES

Win K.M., Kham N.S.M.
University of Computer Studies, Yangon, Myanmar
e-mail: winn.km05@gmail.com, moonkhamucsy@gmail.com

Abstract: Nowadays, microarray datasets are characterized by a large number of gene expression levels for each patient and a relatively small number of patients. Data can grow along two dimensions of fields (called attributes) and the number of cases, abnormal functionalities. Early detection for possible cancer cannot get by large scale data. The problem become feature selection with classification accuracy because of small samples with high dimensionality of genes .The best choice of gene subset means selection of relevant features that is a key for building a more accurate classifier.

The traditional statistical methods with linear function are not capable of discovering nonlinear relationships in microarray data. Data overfitting arises when the number of features is very large. Identifying relevant variables give more insight into the nature of corresponding classification problem and tend to be better predictive performance. Incorporating feature selection is a fairly straightforward procedure for linear or nonlinear support vector machine (SVM) classifiers.

We propose a new method that uses support vector machines with features ranking technique based on recursive feature elimination (RFE).Among various feature selection methods, embedded approach of feature ranking is particularly attractive. A fixed number of top ranked features are selected for design of classifier; a threshold can be set on the ranking criterion. The feature with the smallest ranking criterion is eliminated at each step. The ranking criterion is obtained from the weights of SVM that trained on the subset of features. SVM composes the transformation function and the dot product in the higher dimensional space into a single kernel function. The goal of maximum margin classification in binary SVM is to separate the two classes by a hyperplane. We can decide the optimal hyperplane which reduces the generalization error. The main objectives are to obtain profiles of relevant genes tend to predict the cancer class of unknown patient, prediction accuracy may improve by discarding irrelevant variables and apply the proposed method into an interface that can easily be used by clinicians.

This work is to be optimized which SVM model selection will be particularly important and scalable in microarray data analysis. From our work, experiments on real data sets from UCI machine learning repository can indicate that new method outperforms other feature selection method in terms of classification accuracy. On the other hand, machine learning methods may be able to objectively interpret all available results for the same patient and increase the diagnostic accuracy for each disease.

ANDCELL: A COMPUTER SYSTEM FOR AUTOMATED EXTRACTION OF KNOWLEDGE ABOUT MOLECULAR GENETIC INTERACTIONS AND REGULATIONS FROM PUBMED ABSTRACTS AND THEIR REPRESENTATION AS SEMANTIC ASSOCIATION NETWORKS

Yarkova E.E.*, Demenkov P.S., Ivanisenko V.A.
Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
Novosibirsk State University, Novosibirsk, Russia
Sobolev Institute of Mathematics SB RAS, Novosibirsk, Russia
e-mail: aman@bionet.nsc.ru
* Corresponding author

Motivation and Aim: Work with scientific literature is required for research in every knowledge area. The PubMed database currently contains about 15000000 of scientific abstracts. Their number increases annually by 1 million of abstracts. Analysis of this bulk of literature data, including search of sources, definition of relationships among the described facts is much time-consuming. Furthermore, the modern approaches to analysis of the literature data make obligatory reference to another important source, the factographic databases for molecular biology and genetics.

Methods and Algorithms: The method for automated extraction of information about molecular-genetic interactions from PubMed abstracts was developed using the text-mining approach. For text-mining, we used the previously developed thesauruses for the names of proteins, genes, microRNAs, metabolites, biological pathways, diseases, cells, and organisms. To recognize facts describing molecular genetic interactions in abstract texts we created more than 4000 patterns or decision rules. Extracted information was integrated through reconstruction of networks for semantic associations joining literary facts about molecular-genetics regulations, physical interactions, also about associations between molecular-genetic objects, biological processes and diseases.

Results: The ANDCell system contains the knowledge base and the ANDVisio program for associative network reconstruction. The ANDCell knowledge base contains about 5 millions of molecular genetic interaction facts. The ANDVisio program allows user to access the database and represents the results in a graphic form as associative networks. The vertices of such networks are molecular genetic objects, diseases and processes while the edges between the vertices represent types of relations. The system is provided with a user's friendly interface implemented links to the molecular-genetic databases and articles from information was extracted.

Conclusion: The computer system for automated extraction of knowledge from PubMed abstracts and databases about molecular genetic interactions, gene regulations, catalytic processes, polymorphism gene - disease associations and other associations between facts and their representation as semantic association networks was developed. The ANDCell system may be useful for resolving a wide range of tasks in biology and biomedicine.

Availability: The ANDCell system soon will be available on ICG web server.

Work was supported in part by RFBR: 08-04-91313-IND_a, state contract FASI №02.514.11.4065, interdisciplinary integrative project for basic research of the SB RAS № 115 and RAS presidium program “Molecular and cellular biology”, the grant “Systems biology: computer and experimental approaches.

THE MODEL OF TRANSFERRIN UPTAKE BY CELL: A NOVEL MODE OF TFR2-MEDIATED IRON SEQUESTRATION IN OXIDATIVE STRESS

Yevshin I.S.[1, 2], Sharipov R.N.[1, 2, 3], Shatalin Yu.V.[4], Naumov A.A.[4], Ermakov G.L.[4], Potselueva M.M.[4], Sukhomlin T.K.[4]

[1] Institute of Systems Biology, Novosibirsk, Russia;
[2]Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia;
[3]Design Technological Institute of Digital Techniques, SB RAS, Novosibirsk, Russia;
[4] Institute of Theoretical and Experimental Biophysics RAS, Pushchino, Russia.
* e-mail: ivan@systemsbiology.ru

Motivation and Aim: Iron is a vital ion supporting cell proliferation, metabolism and many cell-specific functions. Oxidative stress (OS) and inflammation result in increase of levels of both non-transferrin bound iron (NTBI) and transferrin (Tf)-bound iron (TBI). TBI is not available for ROS and, thus, does not catalyze HO• production. Cells express Tf receptors, TfR1 and TfR2, to uptake TBI [1]. Apparently, Tf level decreases in OS, although there is no proper explanation of the phenomenon to date. We hypothesize that in OS uptake of diferric transferrin ($Tf:Fe_2$) by the cells specialized to sequester and/or to transfer iron is enhanced due to specific TfR2 properties. For instance, Tf concentration may fall down due to biphasic dependence on $Tf:Fe_2$ uptake by TfR2(+) cells. The increased Tf saturation results both from increase of NTBI in OS and from TfR2-mediated Tf uptake and, therefore, it enhances TfR1-mediated endocytosis, because of TfR1 higher affinity to $Tf:Fe_2$ versus Tf and Tf:Fe. Increased Tf saturation further activates TfR2-mediated $Tf:Fe_2$ uptake due to TfR2 stabilization. So, both iron and Tf uptake should be enhanced in OS. Recently, we have observed a non-linear pattern of Tf level decrease in plasma of the rats with transplanted ascitic tumour [2]. In line with our hypothesis, TfR2 main function is to regulate extracellular iron content rather than just to be OS sensor, as it is currently accepted. The main goal of the work was to construct a model of iron and Tf uptake by the TfR2(+) cells in OS and to test the hypothesis about the role of TfR2 in iron regulation.

Methods and Algorithms: The model of TfRs-mediated iron uptake was constructed in terms of chemical kinetics. The system of ordinary differential equations was solved numerically using BioUML workbench (http://www.biouml.org). The values of model parameters were obtained from available literature. All collected data were deposited in the BMOND database (http://bmond.biouml.org).

Results: Dynamics of $Tf:Fe_2$ in plasma was simulated on the basis of time course of total iron experimentally obtained in OS. TfR2 dynamics resulted from increase of Tf saturation was simulated. Time courses of extra- and intracellular Tf were obtained and fitted to the experimental kinetic curves. Suggested model illustrates observed dynamics of blood Tf and is able to predict critical changes in blood resulted from iron overload in OS. The model may be applied to development of the methodology of OS clinical monitoring in patients with systemic diseases.

Availability: The model is available in BMOND at http://bmond.biouml.org.

References:

1. H. Kawabata et al. (1999) Molecular cloning of transferrin receptor 2. A new member of the transferrin receptor-like family, *J Biol Chem*, **274**: 20826-32.
2. Yu.V. Shatalin et al. (2008) Differential change of ceruloplasmin and transferrin in plasma and ascitic fluid of tumor bearer, *Siberian concilium,* **2**: 59-65.

NEW AND OLD *HOBO* SEQUENCES ARE DIFFERENTLY DISTRIBUTED IN THE GENOME OF *DROSOPHILA MELANOGASTER* STRAIN *Y CN BW SP*

Zakharenko L.P.*, Perepelkina M.P.
Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
e-mail: zakharlp@bionet.nsc.ru
* Corresponding author

Motivation and Aim: Transposable elements (TE) constitute a considerable part in many eukaryotic genomes, influence both the work of genes and the evolution of genomes; therefore, the interest to TE distribution patterns in genomes is well understandable. Usually, the total TE distribution patterns are analyzed; these patterns are determined by two processes—TE insertion and subsequent excision. Presumably, this is the reason why analysis of the same patterns leads to controversial results. We have analyzed the distribution of *hobo* transposon in *Drosophila melanogaster* strain *y cn bw sp,* whose genome is nearly completely sequenced.

Methods and Algorithms: The data from the FlyBase database were used in this work. The sequences were compared using Blast Two Sequences. The sequences were searched for according to specified coordinates with the help of NCBI Mapview (Download Sequence Region *Drosophila melanogaster* (Build 5.1).

Results and Conclusions: Based on experimental and *in silico* data, we have divided the *hobo* into "old" and "new" and demonstrated that the old *hobo* sites are short variable sequences predominantly located near the chromocenter, whereas the predominant location of the new sites in the genome is in the central part of chromosomes. As a rule, the terminal repeats and the central part of DNA sequence, rich with TATA boxes, are lost in the defect *hobo* copies. The only one variant of defect *hobo* that retained its terminal repeats is capable of migrating in this genome. The active *hobo* variants, capable of transposing, are excised, whereas the passive variants are accumulated in the pericentromeric regions. Thus, the inactive defective sequences, presumably, destructed mainly as a result of ectopic recombinations, are accumulated in the chromosome regions with decreased recombination frequency, whereas the active copies transposed mainly by transposases do not follow this pattern.

This work was supported by the Russian Foundation for Basic Research 06-04-48116 and for basic Research of the Presidium of the Russian Academy of Sciences "Dynamics of Plant, Animals and Human Gene Pools".

AGNK: COMPUTER SYSTEM FOR AGNS DATA ANALYSIS

***Zalevsky E.M.*[1, 2 *]*, Mironova V.V.*[1]*, Podkolodnyy N.L.*[1, 2, 3]*, Omelyanchuk N.A.*[1]**
[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
[2] Novosibirsk State University, Novosibirsk, Russia
[3] Institute of Computational Mathematics and Mathematical Geophysics, SB RAS, Novosibirsk, Russia
e-mail: zalevsky@bionet.nsc.ru
[*] Corresponding author

Motivation and Aim: AGNS (Arabidopsis GeneNet Supplementary Database) consists of the two main parts: AGNS ED (Expression Database) and AGNS PD (Phenotype Database) [1]. AGNS ED accumulates gene expression patterns in the wild type, mutants, and transgenic plants. The ED informational unit for the wild type contains the following fields: gene, experiment, developmental stage, anatomical element, expression level and its changes, comments. AGNS PD describes information on phenotypic abnormalities in mutant and transgenic plants. The PD informational unit consists of the following fields: alleles, developmental stage, anatomical element, phenotype anomaly, comments. Comparison of these data sets may allow further systematization of the information and extracting new knowledge. AGNK (Arabidopsis GeneNet Knowledge Database) repository was developed for this aim.

Methods and Algorithms: The special script of data processing was developed for AGNS data analysis. It works out in three following steps. As the first step, the regulatory events in gene expression are extracted using specific query to AGNS ED. The fact of changes in expression level is identified by the presence of the key words in corresponded field. The key words are «switch on», «switch off», «increased» and «decreased». If gene expression is annotated as "present" but not recorded at the previous developmental stages, this entry is also taken but with the special mark. The regulatory events are collected in the XML table 1 and described by the following fields: gene, anatomical element, developmental stage, regulatory event and specific marks. In the second step, information on phenotype abnormalities is extracted for all genes from the table 1. The table 2 integrating these data has the following fields: gene, allele, anatomical element, developmental stage, phenotype anomaly and specific separate marks for gain-of-function and loss-of-function mutations. In the third step, the program shows the fields with the same name in both tables for expert assessment through pairwise comparisons. For example, the anatomical elements may be considered as equal if their relationships are "developed from" or "is a part of".

Results: The final AGNK table contains the following fields: gene, allele, anatomical element, developmental stage (table 1), regulatory event, developmental stage (table 2), phenotype anomaly and specific marks. This XML format table contains cause-and-effect processes in gene networks related to specific developmental abnormalities. The special applet was developed to visualize AGNK table using Java technology.

Conclusion: AGNK analyzes the AGNS data to present the regulatory events in gene expression preceding and predetermining the phenotype abnormality in mutant and transgenic plants. AGNK database is filled up automatically as AGNS data increased.

Availability: {http://wwwmgs.bionet.nsc.ru/agns/agnkapplet/}.

References:

1. N. Omelyanchuk et al. (2006) AGNS - a database on expression of arabidopsis genes, In: *Bioinformatics of Genome Regulation and Structure II.* Eds. N. Kolchanov, R. Hofestaedt, L. Milanesi, 433-442 (Springer Science+Business Media, Inc).

DETECTING CONSERVED WATER MOLECULES IN PROTEIN-DNA COMPLEXES BY COMPARATIVE ANALYSIS OF X–RAY STRUCTURES

Zanegina O.N.*[1], Aksianov E.A.[2], Alexeevski A.V.[2], Karyagina A.S.[3, 4], Spirin S.A.[2]
[1] Bioengineering and Bioinformatics Faculty, Moscow State University, Moscow, Russia
[2] A.N. Belozersky Institute, Moscow State University, Moscow, Russia
[3] N.F. Gamaleya Research Institute of Epidemiology and Microbiology, Moscow, Russia
[4] Institute of Agricultural Biotechnology, Moscow, Russia
e-mail: zanolya@ya.ru
* Corresponding author

Motivation and Aim: It is known that water molecules are involved in protein-nucleic acid interactions. They form specific and non-specific "water bridges" between molecules of NA and protein [1]. The aim of the present work is to find functionally important water molecules by a comparative analysis of X-ray structures of related DNA-protein complexes.

Methods and Algorithms: Automatic tool *wLake* developed in our laboratory [2] was used to find "conserved water molecules" (ConWMs), i.e., those located in almost the same positions in a set of superimposed structures of related proteins or macromolecular complexes. Briefly, the algorithm performs all pairwise superimposition of the structures and detects all overlapped water molecules. Then sets of molecules from the different structures overlapping each other are found, and after a special statistical test some of those sets are detected as ConWMs. The program is available online at http://monkey.belozersky.msu.ru/~evgeniy/wLake/wLake.html).

Results: Using wLake, 84 SCOP domain families containing more than one structure of protein in contact with DNA and water molecules in PDB files were analyzed. All structures described in those PDB files are proved to contain water molecules on DNA-protein interface. ConWMs were found for 80 of families. Structures from 4 remaining families could not be well superimposed so ConWMs were not found. Relation between the number of ConWMs for each family and the number of superimposed structures could not be revealed, because of difference in numbers of water molecules in superimposed structures. We investigate a dependence of number and functional importance of ConWMs with respect to the type of protein-DNA interaction. All analyzed SCOP domain families were classified by the type of protein-DNA interaction based on recognizing elements of protein (helix, sheet, strand, turn) and DNA (the major/minor groove, sugar-phosphate backbone). In most cases ConWMs found in families with several protein recognizing elements were not strongly specific and are immobilized on protein–sugar-phosphate backbone. On the contrary, in recognition of DNA by a single protein element ConWMs are usually more specific and play important role. Some water molecules forming the bridges can present on the protein or DNA surface before the complex is formed. Our approach allows investigating the preimmobilization of the water molecules forming the bridges on the protein or DNA surface. For this aim we (1) compare the detected ConWMs on the protein-DNA interfaces with the known sites of DNA hydration and (2) compare the interfacial ConWMs with ConWMs detected on the surface of the same domains solved without DNA.

Acknowledgements: the work is supported by RFBR, grants 06–04–49558 and 06–07–89143, and INTAS, grant 05–1000008–8028.

References:

2. A.Karyagina et al. (2005). The role of water in homeodomain-DNA interaction. In N.Kolchanov and R.Hofestaedt, (eds), *Bioinformatics of Genome Regulation and Structure II.*, Springer Science+Business Media. pp. 247–257.
3. E. Aksianov, A. Grishin, O. Zanegina, S. Spirin, A. Karyagyna, A. Alexeevski (2008). Conserved water molecules in X-ray structures. Highlight the role of water in intra- and intermolecular interactions. *Journal of Bioinformatics and Computational Biology,* in press.

NANOCOMPLEXES OF MODIFIED OLIGONUCLEOTIDES AS A NEW APPROACH TO OLIGONUCLEOTIDE DELIVERY

Zenkova M.A., Vlassov V.V.
Institute of Chemical Biology and Fundamental Medicine, SB RAS, 8, Lavrentiev ave., Novosibirsk 630090, Russia
e-mail: marzen@niboch.nsc.ru

Oligonucleotide-based therapeutics represent a promising tool for silencing disease-causing genes, particularly those that encode so-called "non-druggable" targets that are not amenable to conventional therapeutics such as small molecules, proteins, or monoclonal antibodies. However, poor uptake of oligonucleotides by target cells remains to be the main obstacle significantly complicating their *in vivo* implementation. Oligonucleotides penetration through the cellular membrane can be improved by association of oligonucleotide molecules into supramolecular complexes and their accumulation on the cell surface. We demonstrated that formation of supramolecular structures by the oligonucleotides enhances their ability to bind with several cancer cell lines. Chemically modified supramolecular complexes, bearing cholesterol on the sense strand, are efficiently taken up by cells. Studies on cellular distribution and biological activity of the complexes formed by delivered oligonucleotides and cholesterol-modified second strand oligonucleotides revealed that these complexes distribute almost uniformly in the cells cytoplasm, they are non-cytotoxic and do not cause any unspecific effects. We demonstrated that incorporation of the desired antisense oligonucleotide into the self-assembling supramolecular system significantly promotes its penetration through the cellular membrane does not required supplementary transfection agents and provides specific inhibition of the target gene.

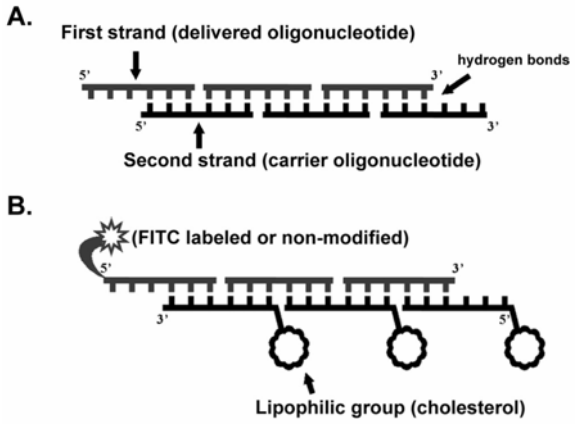


Fig. 1. Types of supramolecular oligonucleotide complexes:
(A) non-modified concatemeric complex;
(B) concatemer, modified with cholesterol and FITC.

This work was supported by Russian Academy of Sciences (Programs "Molecular and Cellular Biology" and "Sciences to Medicine") and SB RAS Interdisciplinary grant №20, RFBR 08-04-00753a/

References:

2. Simonova O.N., Vladimirova A.V., Zenkova M.A., Vlassov V.V. (2006) Enhanced cellular of concatemeric oligonucleotide complexes. *Biochem. Biophys. Acta*, **1758**, 413-418.
3. Simonova O.N., Pyshnyi D.V., Vlassov V.V. and Zenkova M.A. (2008) Modified concatemeric oligonucleotide complexes: new system for efficient oligonucleotide transfer into mammalian cells. *Gene Therapy*, in press.

CLASSIFICATION AND FUNCTIONAL CHARACTERIZATION OF THE HECT-DOMAIN UBIQUITIN-PROTEIN LIGASES

Zhabereva A.S.*, Chaplygina E.V., Okunev O.E., Gainullin M.R.
Nizhny Novgorod State Medical Academy, Nizhny Novgorod, Russia
e-mail: anastasia@gma.nnov.ru
* Corresponding author

Motivation and Aim: The superfamily of ubiquitin-protein ligases (EC 6.3.2.-) plays a key role in the final attaching of ubiquitin to target proteins. There are three E3 ligases families known as of today: HECT, RING and U-box. Whereas RING and U-box proteins serve as scaffolds for ubiquitin transfer from E2 to a target protein, HECT proteins form a thioesther with ubiquitin by an active cysteine residue before transferring it to a substrate. We have done a complex bioinformatic study aimed at development of HECT domain E3 ligase classification. For this purpose, we have analyzed phylogeny, structure-function relationship, protein-protein interactions and human protein expression profiles.

Methods and Algorithms: From PROSITE database we queried HECT domain ubiquitin-protein ligases from *H.sapiens, M.musculus, D.melanogaster, A.thaliana, C. Elegans, S.pombe* and *S.cerevisiae*. The total number of proteins was 138. The search for homologous proteins was performed using the BLAST algorithm. The pairwise and multiple sequence alignments were prepared using the BioEdit and ClustalW programs. The phylogenetic trees were generated using the neighbor-joining algorithm of ClustalW and PHYLIP (distance method), with correction for multiple substitutions, and 1000 bootstrap calculations. The analysis of protein domain organization was executed by NCBI Conserved Domain Database. The search of human protein expression profiles was executed by NCBI UniGene Database transcriptome libraries. For the search of data on protein interaction, BioGRID BETA was used.

Results: The analysis of homology and evolutionary relationship inside dataset allowed us to develop the new classification of HECT domain E3 ligases. All examined proteins are divided into nine subfamilies. Every subfamily is characterized by the structural similarity. The subfamilies are arranged into 3 phylogenetic groups with the common evolutionary origin. Domain analysis of HECT-domain ligases proved the phylogenetic data. Besides the presence of HECT domain, every subfamily possesses its own unique set of domains. Analysis of protein-protein interactions and functions of HECT-domain ligases allowed us to functionally describe every subfamily. Functional analysis allowed a wide variety of processes in which the following HECT domain ligases participate: proteasomal degradation, regulation of transcriptional activity, control on the receptor functions, membrane trafficking. Besides, for most HECT-domain ligases a phenomenon of autoubiquitylation is typical. This variety of biological effects once again underlines the importance of ubiquitin system for cell and shows that the switching off of any component of the system will inevitably result in the development of pathological processes. Particularly, in our research we used the transcriptome analysis to predict the possible role of HECT-domain ubiquitin-protein ligases in cancerogenesis. Between the chosen pathologies (adrenal tumor, cervical tumor, esophageal tumor, kidney tumor, liver tumor, breast tumor, ovarian tumor and skin tumor) a marked activation or repression of expression can be shown for most of HECT-domain ubiquitin-protein ligases.

Conclusion: The results, obtained in present study, allows us to predict the possible functions for the new uncharacterized proteins of HECT-domain E3 ligase family.

AN USING OF DL-SYSTEMS TO MODEL OF THE RENEWABLE ZONE SIZE CONTROL IN GROWING TISSUE

***Zubairova U.S.*[1*]*, Nikolaev S.V.*[1]**
[1] Institute of Cytology and Genetics, SB RAS, pr. Lavrentieva 10, Novosibirsk, 630090, Russia
e-mail: ulyanochka@gorodok.net
* Corresponding author

Motivation and Aim: The shoot apical meristem is a small group of nondifferentiated dividing cells that generate all of the aerial parts of the plant. The meristem structure remains constant during plant growth, but its resident cells changes. As a result of horizontal division of central zone cells, the cells shift down and transform into cells of organizing center. In turn the cells of organizing center also shift down and transform into the cells of rib-zone. In our work we considered cell growth, and stochastic cell divisions depending on their size. Cells of shoot apical meristem are determined to express certain genes. 2-4 cells around vertical axis of meristem in 3-4 uppermost layers express CLV3. It is the central zone. The cells beneath the central zone express WUS. Mechanism that provide such constant structure is the subject of intensive research.

Results: The main concept of our model is the following. We observed one dimensional array of cells on the vertical axis of meristem. Substances Y, Z, W can be synthesized in the cells with rates depending on the concentrations of these substances. It is assumed that Y and W are diffusible. The Y is synthesized in first cell and diffuses through the cells of the array. The rate of the Y synthesis depends on concentration of W in the cell. The C does not diffuse and only decay. The rate of the C synthesis depends on concentration of Y. Substance W diffuse through cell-array and regulate synthesis of Y in the first cell. Its rate of synthesis depends on concentrations of Y and C. Thus we can formulate this model in terms of Cauchy problem.

Another part of model is cell division. Production of substance W defines the lower boundary of renewable zone. Each cell is characterized with its length. At initial time all cells are labeled with initial values of length according to normal distribution. Next time cells grow and value of l increase. When the length achieves critical value, the cell divides in certain relation k, which is also stochastic variable. Concentration of substances in its child cells is the same as in the parent cell. Lengths of the cells outside of the renewable zone are unchanged. In this work we studied influence of the cell division on dynamics of compartmental structure of the renewable zone. We also interested in relations between character time of the cell cycle and diffusion of morphogens and influence of this factors on the systems stability.

Model is realized in stochastic parameterized dL-system. The key concept is the integration of discrete and continuous aspects of model behavior into a single formalism, where L-system-style productions express qualitative changes to the model, and differential equations capture continuous processes. We used program package *Mathematica* 5.2 for modeling.

Conclusion: Computer simulations with the model demonstrated movement of the zone boundaries as a result of occasionally simultaneous division of some cells. These movements was followed by recovering “normal” structure. This model behavior is in consistence with some experimental data. However, in some computer experiments we observed destruction of the system, when many cells divide occasionally in the same time.

Acknowledgements: This work was supported by the RFBR grant: 08-04-01214-a "Mathematical modeling and analysis of structure homeostasis mechanism of the stem cell niche in Arabidopsis thaliana shoot apical meristem"

WHOLE-GENOME COMPARISON OF TWO MYCOBACTERIUM TUBERCULOSIS STRAINS BY THE PROGRAM NUCLON 2.0

Zubov I.V.[1], Zubov V.V.[2]*
Vyatka State University, Kirov, Russia[1]
Institute of Theoretical and Experimental Biophysics RAS, Puschino, Russia[2*]
e-mail: genseq@mail.ru
[*] Corresponding author

Motivation and Aim: Immunoprotective epitopes of proteins from chronic infection pathogens, including *Mycobacterium tuberculosis*, are noted for its high mutability. The development of antituberculosis vaccines of new generation requires a search for these epitopes. The purpose of the present work was a comparative genome analysis of two tuberculosis pathogen strains and the detection of hypervariable gene regions using the NUCLON software [1].

Methods and Algorithms: The analysis of genetic variability was carried out using the NUCLON 2.0 program with a simplified algorithm of Nei-Gojobori [2] and a built-in MEGABLASTN module. The annotated nucleotide genome sequences of *M. tuberculosis* CDC155 and H37Rv were compared.

Results: Only 49 of 4187 CDS of the strain CDC155 have more than 5 mutations. Considerable genetic variability is typical for 15 PE_PGRS and 8 PPE family proteins. Local regions with multiple nucleotide substitutions appear in several unstudied (hypothetical proteins 722, 371, 239, 110 AA) and poorly studied proteins (glycosyl hydrolase, 1254 AA; primosomal protein N, 655 AA; alkyl-dihydroxyacetonephosphate synthase, 433 AA; alpha/beta hydrolase, 293 AA; RNA polymerase sigma-70 factor, 168 AA). The greatest amount of mutations is found in adenylate cyclase (531 AA, 245 mutations) and phospholipase C (519 AA, 433 mutations). The most clearly pronounced local hypervariable regions were detected in the sugar transporter family protein (506 AA) and lipoprotein (219 AA).

Conclusion: The results obtained are largely consistent with earlier published data [3] and show that the program NUCLON 2.0 is a promising tool for a comparative genome analysis of *M. tuberculosis* and other pathogens.

Availability: available on request from the authors.

References:

1. A.B.Dolzhenkov, V.V.Zubov (1997) Principles of the search for immunogenic epitopes, In: *Theses of reports of the conference devoted to the 100 anniversary of the foundation of the Russian antiplague service (16—18 September, 1997, Saratov)*, **1:** 251.
2. M.Nei, T.Gojobori (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions, *Mol. Biol. and Evolution*, **3:** 418-426.
3. R.D.Fleischmann et. al. (2002) Whole-genome comparison of Mycobacterium tuberculosis clinical and laboratory strains, *J. Bacteriol*., **184:** 5479–5490.

AUTHOR INDEX

D

E

F

G

H

I

J

K

L

M

N

O

P

Q

R

S

T

U

V

W

X

Y

Z

Научное издание

Труды шестой международной конференции
“Биоинформатика регуляции и структуры генома”

на английском языке

Proceedings of the Sixth International Conference
on Bioinformatics of Genome Regulation and Structure

Abstracts have been printed without editing
as received from the authors

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

Дизайн и компьютерная верстка: А.В. Харкевич

Подписано к печати 26. 05. 2008 г.
Формат бумаги 70×108 1/16. Печ. л. 24,2. Уч.-изд. л. 31,1
Тираж 250. Заказ 167

Отпечатано в типографии Издательства СО РАН
630090, Новосибирск, Морской пр., 2