Moscow State University
Institute for Information Transmission Problems RAS
N.F.Gamaleya Research Institute of Epidemiology and Microbiology RAMS
Vavilov Institute of General Genetics RAS
Moscow Institute of Physics and Technology
Russian Foundation for Basic Research
Skolkovo Foundation
German Research Foundation

P R O C E E D I N G S

OF THE INTERNATIONAL MOSCOW CONFERENCE ON COMPUTATIONAL MOLECULAR BIOLOGY

MCCMB'11

Moscow, Russia,
July 21-24, 2011

Organizers

Department of Bioengineering and Bioinformatics of M.V. Lomonosov Moscow State University
www.fbb.msu.ru

Biological Department of
M.V. Lomonosov Moscow State University
www.bio.msu.ru

Institute for Information Transmission Problems of the Russian Academy of Sciences (Kharkevich Institute)
www.iitp.ru

N.F.Gamaleya Research Institute of Epidemiology and Microbiology,
Russian Academy of Medical Sciences
www.gamaleya.ru

Moscow Institute of Physics and Technology (State University)
phystech.edu/about/

Vavilov Institute of General Genetics,
Russian Academy of Sciences
www.genetika.ru

Russian Foundation for Basic Research
http://www.rfbr.ru

German Research Foundation
http://www.dfg.de/

Skolkovo Foundation
http://www.i-gorod.com

Organizing Committee

Vladimir Skulachev
Faculty of Bioengineering and Bioinformatics of the MSU, Moscow, Russia |
co-chair

Mikhail Kirpichnikov
Biological Faculty of the MSU, Moscow, Russia
co-chair

Aleksander Kuleshov
Kharkevich Institute for Information Transmission Problems, RAS, Moscow, Russia
co-chair

Vladimir Tumanyan
Biophysics Council of RAS, Moscow, Russia

Mikhail Gelfand
Kharkevich Institute for Information Transmission Problems RAS, Russia
chair of the program committee

Vsevolod Makeev
Vavilov Institute of General Genetics, Russia
deputy chair

Mireille Regnier
INRIA, France

Shekhar Mande
University of Hyderabad, India

Dmitry Frishman
Technical University of Munich, Germany

Andrei Alekseevsky
Belozersky Institute of Physical-Chemical Biology, MSU, Russia

Sergei Spirin
Belozersky Institute of Physical-Chemical Biology, MSU, Russia

Sophia Rodionova
Kharkevich Institute for Information Transmission Problems, RAS, Russia
secretary

Program Committee

Georgy Bazykin
Kharkevich Institute for Information Transmission Problems, RAS, Russia

Alexey Finkelshtein
Institute of Protein Research RAS, Russia

Dmitry Frishman
Technical University of Munich, Germany

Mikhail Gelfand (Chair)
Kharkevich Institute for Information Transmission Problems, RAS, Russia

Anna Karyagina
N.F. Gamaleya Research Institute of Epidemiology and Microbiology, RAMS, Russia

Alexei Kondrashov
University of Michigan, United States

Vsevolod Makeev
Vavilov Institute of General Genetics, Russia

Yuri Panchin
Kharkevich Institute for Information Transmission Problems RAS, Russia

Vladimir Poroikov
Institute of Biomedical Chemistry of RAMS, Russia

Mireille Regnier
INRIA, France

Mikhail Roytberg
Institute of Mathematical Problems of Biology, Russia

CONTENTS

Statistical analysis of errors' context for Illumina sequencing

Irina Abnizova[1], Rene te Boekhorst[2], Steven Leonard[1], Tom Skelly[1], Tony Cox[1]

[1]*Wellcome Trust Sanger Institute, United Kingdom, ia1@sanger.ac.uk*
[2]*University of Hertfordshire, United Kingdom.*

The new generation of short read sequencing technologies still requires both accuracy of data processing methods and reliable measures of that accuracy and data quality. Such measures are especially important for variant calling. However, in the particular case of SNP calling, a great number of False Positive SNPs (FP SNP) may be obtained. Reliable methods are required to distinguish putative SNPs from sequencing- or other errors.

We found that not only the probability of sequencing errors (i.e. the quality value) is important to distinguish a FP SNP, but also the conditional probability of 'correcting' this error (the 'second best call' probability, conditional on that of the first call). Surprisingly, around 80% of mismatches can be 'corrected' with this second call.

Another way to reduce the rate of FP SNPs is to retrieve DNA motifs which seem to be prone to sequencing errors, and to attach a corresponding conditional quality value to these motifs. We have developed several measures to distinguish between sequence errors and candidate SNPs, based on a base call's nucleotide context and its mismatch type.

In addition, we suggested a simple method to correct the majority of mismatches, based on conditional probability of their 'second' best intensity call. We attach a corresponding second call confidence (quality value) of being corrected to each mismatch.

The software is available from the authors by request. It collects mismatches from a BAM or qseq file, performs statistical analysis of these mismatches and assigns probability of being corrected by the second intensity call to each mismatch.

In Illumina technology [1], the sequencer analyses one nucleotide of the sequence in each cluster per cycle. At each cycle, A,C,G and T nucleotides, each labeled with a different dye, are added to the flow-cell and an intensity for each dye in each cluster is recorded. Ideally, the strongest of the four intensities recorded at a given cycle for a given cluster should correspond to the nucleotide at that position in the sequence for that cluster. Two problems accrue in clusters of DNA over successive cycles of the Illumina sequencer: phase inaccuracy due to base-incorporation errors and dye-label cross-talk, effectively limit the use at this technology to short read lengths of around 100 bases on the GA2 and HiSeq.

There are attempts to solve these problems: Alta-Cyclic [2] applies a machine learning method to improve a base call; AYB [3] takes an account of DNA context to improve a base call. Statistical models for phase inaccuracy, drop-off and dye-label cross-talk are combined by W. Gilks [4] in a base-calling inference tool, which allows for differential rates of phase inaccuracy and drop-off between DNA clusters, and calculates the probability of miscall for each called base. Other methods try to account for nucleotide context by introducing such a predictors for quality values as di-nucleotide content [5] or homo-polymer count [1]. However, all these methods do not analyze possible errors directly.

Let us briefly clarify the notation we will use further. Any mismatch is defined by the nucleotide called and the corresponding nucleotide in a reference genome; this reference nucleotide is obtained after mapping called read to the reference genome. There are twelve possible mismatch types of substitutions: A-> C, A-> G, A-> T, C-> A, C-> G, C-> T, G -> A, G -> C, G-> T, T->A, T-> C, and T-> G. The first letter for each pair is what it is in a reference, and the second letter stands for what was called by sequencing. When we say 'mismatch pattern', we think about an identity of a base called (as expected in a reference), and an identity of its preceding base(s), namely about DNA context preceding a mismatch position.

While looking at the mismatches we noticed some intriguing facts:

o Certain mismatch patterns and mismatch types occur significantly more frequently than others, and are common for organisms and runs

These observations inspired us to perform a systematic analysis of Illumina mismatches/errors, and to see which mismatch types and mismatch patterns are persistent between different lanes, runs, organisms, CG-contents and machines for Illumina sequencing. Looking at the mismatch context, we aimed to learn if an identity of previous base(s) affects the mismatch type and frequency.

We also kept in mind the cross-talk and phasing inaccuracy tendencies mentioned before. These error tendencies might be enhanced for particular DNA combinations (local DNA content), making them especially error prone.

1. http://www.illumina.com/Documents/products/technotes/technote_casava_secondaryanalysis.pdf
2. http://www.ebi.ac.uk/goldman-srv/AYB/
3. http://www.slidefinder.net/s/statistical_base_caller_illumina_genome/12171624/p3
4. http://www.broadinstitute.org/gsa/wiki/index.php/Base_quality_score_recalibration
5. Erlich Y, Mitra PP, Delabastide M, et al. Alta-Cyclic: a self-optimizing base caller for next-generation sequencing. NatureMethods 2008;5(8):679–82.

Analysis of the transcriptome of the human parasitic trematode Opisthorchis felineus

Dmitry Afonnikov[1], Mikhail Pomaznoy[1], Alexey Katokhin[1], Vyatcheslav Mordvinov[1], Nikolay Kolchanov[1], Kostryukova E.S.[2], Levitskii S.A.[2], Selezneva O.V.[2], Chukin M.M.[2], Larin A.K.[2], Lazarev V.N.[2], V.M. Govorun[2]

[1]*Institute of Cytology and Genetics SB RAS, 10 Lavrentyev Ave, Novosibirsk, 630090, Russia, ada@bionet.nsc.ru, pomaznoy@gmail.com, katokhin@bionet.nsc.ru, mordvin@bionet.nsc.ru, kol@bionet.nsc.ru*
[2]*Scientific Research Institute of Physical-Chemical Medicine, 1a Malaya Pirogovskaya Str., Moscow, 119992, Russia, lazar0@mail.ru, govorun@hotmail.ru*

The liver fluke *Opisthorchis felineus* cause the opisthorchiasis and represents a substantial public health problem in Siberia and eastern regions of the former USSR. From 40000 to 95000 cases of opisthorchiasis were reported annually between 1986 and 1992 [1]. In the central part of Western Siberia, the prevalence ranges between 40% and 95%. The opisthorchiasis is associated with a number of hepatobiliary abnormalities, including cholangitis, obstructive jaundice, hepatomegaly, cholecystitis, cholelithiasis and carcinogenesis [2]. Therefore, it is important to search for methods of efficient combating with *O. felineus* infection, based on a detailed knowledge of the interplay between the parasites and their hosts as well as the biology of the parasites themselves at the molecular level.

The life cycles of *O. felineus* involves an aquatic snail, in which asexual reproduction takes place, and freshwater fishes as intermediate hosts. Fish–eating mammals, including humans, dogs and cats, act as definitive hosts, in which sexual reproduction occurs [3].

In this work we performed the bioinformatics analysis of transcriptome sequences from *O. felineus* marita (adult form parasitizing in mammals liver). As result, structural and functional annotations of more than 20000 sequences were obtained.

In our report we will present the results of analysis of structure and functional characteristics of the *O. felineus* transcriptome for this life form.

The work was supported by RAS programs № 6.8, Б 26.29 и 24.2., SB RAS integration projects 113 and 119, the SB RAS Program "Genomics, Proteomic, Bioinformatics", the State contract Rosnauki 02.512.11.2332, the RFBR grant 09-04-12209-ofi_m.

1. WHO study group on the control of the foodborne trematode infections. (1995) Control of the foodborne trematode infections: report of a WHO study group, *Who technical report series*, **849:**92–93.
2. Mordvinov VA, Furman DP. (2010) The Digenea parasite Opisthorchis felineus: a target for the discovery and development of novel drugs. *Infect Disord Drug Targets*. **10**:385-401.
3. Beer SA. (2005) Biology of opisthorchosis invasion agent. Moscow: KMK Press;

Lack of dominance effect revealed by comparison of frequencies of nonsense mutations between autosomes and X-chromosome in Drosophila melanogaster

Aleksandra Akhmadullina[1], Georgii Bazykin[2], Alexey Kondrashov[3]

[1]*Moscow State University, Russian Federation, aleksandra.akhmadullina@gmail.com*
[2]*Institute for Information Transmition Problems, Russian Federation, gbazykin@iitp.ru*
[3]*University of Michigan, United States, kondrash@umich.edu*

Despite decades of debate, our understanding of dominance effects remains elusive. Since the seminal works of Fisher and Wright, it has been assumed that large-effect mutations tend to be recessive, while the effects of small-effect mutations is usually additive. Nonsense mutations prevent the synthesis of the gene product, and can be safely assumed to be large-effect; nevertheless, they can reach a substantial frequency in the population. Here, we use 162 complete genome sequences of D. melanogaster to study the frequency of nonsense mutations segregating in the population. In 10.2% of genes in our D. melanogaster sample, stop codons segregate at non-trivial frequencies. Since the alleles carrying stop codons on the X chromosome are hemizygous and more visible to selection, we expected the nonsense mutations to have lower frequencies on the X chromosome. Surprisingly, the fraction of genes carrying a nonsense mutation was virtually identical between the X chromosome (10.0%) and autosomes (10.2%). The allele frequency spectra, which can reveal the action of weak selection, were also similar. The significance of these results for our understanding of dominance is discussed.

1. F.A. Kondrashov, E.V. Koonin (2004) A common framework for understanding the origin of genetic donimamce and evolutionary fates of gene duplications, *TREND in Genetics*, **20:**7.
2. R.A. Fisher (1928) The possible modification of the response f the wild type to recurrent mutations, *Am. Nature*, **62:**115–126.
3. P. Andolfatto et.al (2010) Effective population size and the efficacy of selection on the X chromosomes of two closely related Drosophila species, *Genome Biol. Evol.*, **3:**114–128.

Automatic Detection of Architectures in 3D Protein Structures of All-Beta and Alpha/Beta Classes

Evgeniy Aksianov[1], Andrei Alexeevski[1,2]

1Belozersky Institute for Physical and Chemical Biology, Moscow State University
2 Scientific Research Institute for System Studies (NIISI RAS), Moscow

The classification of protein folds is a hard problem: widely used hierarchical classifications SCOP and CATH vary in significant details, and are based on manual expert decisions for those structures which can't be well superimposed to already classified domains. Attempts to create fully automatic classifications are less successful [1].

Here we present **SheeP** (Sheet Puzzle) program aimed to detect architectures in input protein structure of PDB format. All detectable architectures have core beta-sheets. They are beta-sandwich (parallel and orthogonal), beta-barrel (single), beta-prism, beta-propeller, beta-bigmac (beta-sandwich with more than two layers), alpha/beta barrel, GFP-like barrel, alpha/beta sandwich, Rossmann fold (alpha-beta-alpha sandwich), single sheet (including open barrels), small sheet (considered as non-structural element).

Core procedure of **SheeP** algorithm is beta-sheet detection and description. Graphical output of a beta-sheet is 'a sheet map' (Fig.1). Beta-sheets are detected in several steps. Preliminary beta-sheets are obtained from DSSP program output. Each preliminary beta-sheet undergoes several steps of modifications using geometrical criteria. Modifications may result in breaking preliminary sheet into two separate sheets or, conversely, joining two sheets into one; adding or excluding residues from the sheet and so on. Arrangements of interacting beta-sheets, helixes and beta-sheets, are analyzed to determine architectures. Additionally, standard structural motifs, like jelly-rolls, Greek keys and interlocks, are detected. **SheeP** program is implemented as web-service at http://mouse.belozersky.msu.ru/.

SheeP program was tested on all 125567 domains of CATH DB, release 3.30. Such CATH architectures as beta-sandwich (**SheeP** algorithm detected 92.6% , 14252 of 15395 domains from 38 topologies belonging to sandwich architectures), beta-prism (85.4% ,135 of 158 domains) were detected appropriately well. In CATH, beta-barrel architecture includes single barrels as well as open barrels. Open barrels are determined by **SheeP** as single sheets. This is the main reason why only a half of beta-barrel domains (47.1%, 4500 of 9547) were correctly detected. For example, only 150 of 375 domains from CATH family 2.40.30.10, annotated in CATH as 'barrel, topology: Elongation Factor Tu (Ef-tu); domain 3', were detected

by **SheeP** as closed barrels. Domain 1cqxA151-261 (flavohemoglobin domain) was not detected by **SheeP** as closed barrel and visually it is open barrel; 2b7bA333-440 (elongation factor domain) was detected as closed barrel (and visually, it is). Thus, distinguishing open and close barrels provides additional arguments for the classification. Another problem occurs with corrupt beta-strands and beta-sheets, which are not detected by DSSP algorithm. For example, 4 of 47 "Orthogonal Prisms" domains were not detected correctly by **SheeP** because of lack of one sheet due to a large number of irregularities.

We believe that developed algorithm and further improvements of algorithmic descriptions of architectures and folds could lead to more objective and reproducible fold classifications.

Fig. 1. Sheet map of lectin from Scilla campanulata (PDB code 1dlp, chain B, only one of three sheets is shown). A non-empty cell of the map contains a residue from the sheet, empty cell - 'a gap' - is needed to represent sheet's irregularities. Each strand is contained in one raw. Cross-strand arrays are shaded according to location of side chain: on one side of the sheet, on another side or failed to determine. Bold lines indicate either edge of a strand (N-terminal or C-terminal) or absence of regular hydrogen bonds within a cross-strand array.

The work was partially supported by RFBR grant 10-07-00685-a.

1. Sam V. et al. (2008), Towards an automatic classification of protein structural domains based on structural similarity, *BMC Bioinformatics*, **9**:74

Accurate response timing of a bistable gene switch

Jaroslav Albert, Marianne Rooman
Universite Libre de Bruxelles, CP 165/61, avenue F. D. Roosevelt 50, 1050 Bruxelles, Belgium, jalbert@ulb.ac.be

Switching between two stable states (bistability) of a gene has been observed to be a ubiquitous feature in many organisms. For a system to be bistable, positive autoregulation is required which can be achieved by either a single-gene positive feedback loop [1] or a motif comprising of two mutually suppressing genes [2]. Examples of bistability include: sporulation and competence in bacterium B. subtilis [3] and the maturation of frog oocytes [4]. Studies of the expression dynamics of positively autoregulated genes have shown that the time to reach steady state is always slower than for genes with negative or no feedback [5]. Some functions associated with delays in state transition are: filtering out short-lasting activation signals arising from noise [6]; and, orchestrating temporal chain of events, i. e. cascades, where a set of genes is turned on one by one [7].

Since any biological system is subject to intrinsic noise, one may wonder about the accuracy with which gene switching can be delayed. In this work we set out to answer the following question: what are the necessary conditions, i. e. parameter values, that allow a genetic switch to be delayed with accuracy? In order to answer this question, we considered the simplest bistable system: a single gene with positive autoregulation. As a starting point, we approximated the full system – involving all biochemical reactions -- by replacing the state variables representing the DNA-transcription-factor complexes with their equilibrium values; this allowed us to describe the system with only one second order differential equation corresponding to the protein concentration. With this single equation, the analysis became similar to that of a particle moving in an external potential under a frictional force. Next, we selected many parameter sets, each yielding a different set of three fixed points: metastable (MSP), unstable (UP) and stable (SP). Switching occurs when an external input changes one of the parameter values in such a way that the MSP and the UP merge together, leaving the SP as the only equilibrium solution. If the system initially starts out in the MSP it will be driven to the SP. The time to reach the SP will depend on the amplitude of the external signal.

Finally, returning to the full set of stochastic equations and setting the initial mRNA and protein concentrations (mPC) to the MSP, we simulated, using the Gillespie algorithm [8], the evolution of the mPC under the influence if external input. For every parameter set, the magnitude of the input was chosen whereby the same delay was produced. Our results show that one of the criteria for generating reliable delays is the initial and final (steady state) mPC. At large concentrations, switching delays tend to be more accurate than at low concentrations. However, we also found that for some parameter sets accurate delays can be achieved even at low mPC.

Challenges in Comparative Genomics: from Biological Problems to Combinatorial Algorithms (and back)

Max Alekseyev
University of South Carolina, United States, maxal@cse.sc.edu

Recent large-scale sequencing projects fueled the comparative genomics studies and heightened the need for algorithms to extract valuable information about genetic and phylogenomic variations.

Since the most dramatic genomic changes are caused by genome rearrangement events (that shuffle genomic material), it becomes extremely important to understand their mechanisms and reconstruct the sequence of such events (evolutionary history) between genomes of interest.

In this expository talk I shall describe several controversial and hotly debated topics in evolutionary biology (chromosome breakage models, mammalian phylogenomics, prediction of future rearrangements) and formulate related combinatorial challenges (rearrangement and breakpoint re-use analysis, ancestral genomes reconstruction problem).

I shall further present recent theoretic and algorithmic advances in addressing these challenges and their biological implications.

"Metazoa-specific" genes in the early opisthokonts

Kirill V. Mikhailov, Vladimir V. Aleoshin

Belozersky Institute for Physicochemical Biology, Lomonosov Moscow State University, Moscow, Russia, *Aleshin@genebee.msu.su*
Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow 127994, Russia

The Opisthokonta is a well established group of eukaryotes that includes animals, fungi and choanoflagellates. Phylogenetic analyses have shown that, in addition to its main lineages, Opisthokonta includes a number of lesser known unicellular organisms that were previously classified as diverse Protozoa [1]. The majority of trees based on multigene datasets place these organisms in a monophyletic clade, known as Mesomycetozoea, sister to animals and choanoflagellates. The phylogenetic position of Mesomycetozoea has generated significant interest to the group in the context of emergence of Metazoa [2]. The genome sequencing initiative aimed at these early diverging members of Opisthokonta revealed numerous homologs of genes involved in cell adhesion and signaling that were previously considered specific to Metazoa. The genome of *Capsaspora owczarzaki*, a filose amoeboid member of Mesomycetozoea, contains homologs of metazoan integrins and kinases implicated in focal adhesions [3], and transcription factors that are responsible for cell differentiation in Metazoa [4]. Presence of these genes in members of Mesomycetozoea implies their premetazoan ancestry, and suggests that part of the metazoan genetic toolkit emerged prior to the divergence of metazoan and mesomycetozoean lineages. Surprisingly, many of these genes are absent from the genome of choanoflagellate *Monosiga brevicollis*, which means that they were lost by choanoflagellates.

1. M.A. Ragan, et al. (1996) A novel clade of protistan parasites near the animal-fungal divergence, *Proceedings of the National Academy Sciences of the United States of America*, **93**:11907-11912.
2. I.Ruiz-Trillo, G.Burger, P.W.Holland, N.King, B.F.Lang, et al. (2007) The origins of multicellularity: a multi-taxon genome initiative, *Trends in Genetics*, **23:**113–118.
3. A.Sebé-Pedrósa, et al. (2010) Ancient origin of the integrin-mediated adhesion and signaling machinery, *Proceedings of the National Academy Sciences of the United States of America*, **107**: 10142–10147.
4. A.Sebé-Pedrósa, et al. (2010) Unexpected repertoire of metazoan transcription factors in the unicellular holozoan *Capsaspora owczarzaki*, *Molecular Biology and Evolution*, **28**: 1241-1254.

Protein Sequence Alignments Comparison and Verification

Boris NAGAEV[1], Boris BURKOV[1], Daniil ALEXEYEVSKY[1],
Sergei SPIRIN[1,2,3], Andrei ALEXEEVKI[1,2,3]

[1]*Faculty of Bioengineering and Bioinformatics, Moscow State University*
[2]*Belozersky Institute for Physical and Chemical Biology, Moscow State University,*
[3] *Scientific Research Institute for System Studies (NIISI RAS), Moscow*
aba@belozersky.msu.ru

Here, we report a package MALAKITE, which allows detecting reliable parts in an input multiple alignment of protein sequences, constructed by any other program, and use it for verification and comparison of protein sequence alignments.

Alignment is meant to express an equivalence relation between monomers in sequences. In the case of protein sequence alignments this equivalence can be regarded as residue homology. Standard alignment representation implies homology of all monomers in each column. However, most alignment programs sometimes put unrelated residues in the same column, either due to limitations of algorithms or to aid visual simplicity of representation.

Thus, additional means of detecting and displaying homologous residues in alignment are required. To some extent such means are provided in a number of alignment-related software by coloring schemes (e.g. M-Coffee [1]), or by uppercase/lowercase letters (e.g. Prefab [2]). However, these approaches do not suffice for cases where several classes of homologous residues fall into the same column. We know the only approach, in which residue equivalence (homology) is expressed adequately. It is implemented in POSA service for multiple flexible 3D structure superimposition [3] .

We have developed software to detect residue homology in an input alignment by finding blocks confirmed either by superimopsition of fragments of 3D structures (MALAKITE_3d) or by detectable sequence similarity (MALAKITE_seq). The result of both programs is the decomposition of all residues of all sequences into classes of (presumably) homologous residues; one class is contained in one column, but one column may contain several classes.

In order to verify Pfam alignments, we had to consider each column of Pfam alignment as one class of homologous residues. MALAKITE_3d program was used for verification; thus, only 2223 alignments that contained at least two sequences of proteins with known 3D structure were selected and only sequences with determined 3D structures were left within thus producing reduced alignments.

To measure these reduced alignments' quality we define *number e_i of independent errors in column i* as number of equivalence classes detected by MALAKITE_3d within the column minus one. Average *number E of independent errors per column* is defined as weighted sum $E = \Sigma_i e_i (s_i / \Sigma_j s_j)$ where s_i is a number of residues in column *i*.

We conclude that the majority of Pfam alignments are well confirmed by 3D: 1526 of the reduced alignments have less than *0.5* independent errors per column, although there are 215 alignments with $E>1$ and there are even four alignments with $E>3$.

We generalize value *E* for comparison of any two alignments, *A* and *B*, of the same sequences. Here each alignment is considered as a decomposition of all residues into classes of homologous residues. We define *relative error $E_{(B/A)}$ of alignment B with respect to alignment A* as $E_{(B/A)} = \Sigma_x e_x (s_x / \Sigma_y s_y)$ where *x* runs through classes of *B*, e_x is a number of classes of *A* intersecting *x*, s_x is number of residues in *x*. If we consider *A* more reliable, $E_{(B/A)}$ displays how much B joins separate classes of *A* together and $E_{(A/B)}$ displays how much B splits classes of *A* apart. Value *E* defined above for Pfam alignments verification coincides with $E_{(\text{Pfam/MALAKITE\_3d})}$.

Using value *E* we compared MALAKITE_3d vs MALAKITE_seq for 170 reduced Pfam alignments which have at least 10 sequences and Pfam alignments vs Muscle alignments (the latter showed a bit worse results).

The work is partially supported by RFBR grants 09-04-92743, 11-04-91340, 10-07-00685.

1. S. Moretti at al. (2007), The M-Coffee web server: a meta-method for computing multiple sequence alignments by combining alternative alignment methods. *NAR*, 35:W645-8.
2. R.C.Edgar, (2004), MUSCLE: multiple sequence alignment with high accuracy and high throughput, *NAR*, 32:1792-97.
3. Y.Ye, A.Godzik (2005) Multiple flexible structure alignment using partial order graphs, *Bioinformatics*, 21:2362-2369.

Web-application for comparative structural and functional analysis of prokaryotic genomes sequencing data

Ilya Altukhov[1], Dmitry Ischenko[1], Dmitry Alexeev[2],
Nikolay Bazaleev[2], Alexey Uvarovskiy[1], Alexandr Tyakht[2]

[1]*Moscow Institute of Physics and Technology, Russian Federation, ilya.altukhov@gmail.com*
[2]*Research Institute of Physical Chemical Medicine(Moscow), Russian Federation*

Regarding to widespread using of high-performance technologies of next generation sequencing one of the tasks of bioinformatics is to optimize the algorithmic methods for the analysis of the collected genomic information. Bioinformatics group of Research Institute of Physical Chemical Medicine (Moscow) have created users web-application for interested researchers that provides the most effective software solutions for comparative structural and functional analysis of prokaryotic genomes sequencing data. Comparative analysis consists of an alignment of genomic sequences, the search of single nucleotide polymorphisms with a verifying their synonymy, the genomic annotation based on known genomes structure of the same species of microorganisms, graphic comparison of results. Internet portal enables users to conduct their own genome project, keeping all information in the database, which allows access to analysis results from any PC. The effectiveness of available algorithms is demonstrated in the analysis of structural and functional features of genomes of clinical isolates of Neisseria gonorrhoeae and Mycobacterium tuberculosis, which have been re-sequenced under the scientific research project bearing by institute staffs.

Key Residues in Protein-DNA Interactions

Anastasia Moraleva[1], Eugene Kuznetsov[2], Vladimir Tumanyan[1], Anastasya Anashkina[1]

[1]*Engelhardt Institute of Molecular Biology RAS, Vavilov str., 32 Moscow 11999, Russia*
[2]*Institute of Control Sciences RAS, Prophsoyuznaya str., 65 Moscow, 117997, Russia,*
anastasya.anashkina@gmail.com

Recognition of DNA by proteins is one of the most important processes in living systems. There has been a growing interest in the prediction of DNA-binding sites in proteins which play crucial roles in gene regulation. We have previously developed a method of detection of residues involved in protein-DNA interactions based on Voronoy-Delaunay tessellation [1]. We reported importance of charged residues in DNA recognition both positively charged ARG and HIS and negatively charged ASP and GLU [2].

Here we report the use of two different ways in search of key residues in protein-DNA interactions. Conservation and variation scores are used when evaluating DNA-binding sites in a multiple sequence alignment, in order to identify residues critical for structure or function. Residues involved in protein-DNA contacts are defined by Voronoi-Delaunay tessellation procedure.

All available on 15 January, 2011, about 1900 protein-DNA complexes in PDB [3] were selected for contact analysis. The analysis revealed that amount of residues that formed contact with only one nucleotide is 40%, with two nucleotides is 40%, with three nucleotides is 9% from total amount of interfacial residues. Highest value of contacts finded out for ARG4 from A chain of 1EA4 complex. This arginine placed in major groove of double stranded DNA and formed 17 contacts with different neighboring nucleotides. The medium number of nucleotide neighbors per interfacial residue is 2.0. Minimal value 1.6 belongs to aspartic acid, maximal value 2.7 belongs to arginine.

Quantity of residues formed more then 6 contacts is less 1.3% from total amount of interfacial residues. There are mainly arginines and lysines. One can suppose that these residues (formed contacts with more than 6 different nucleotides through protein-DNA interface) are the most important points in protein-DNA complex formation.We investigated a conservation of these residues (formed 7 and more contacts) by different conservation and variation scores and drew a conclusion that less then 60% of these residues are conservative by Valdar [3] conservation measure.

All revealed residues involved in protein-DNA interactions were used for conservation analysis. Also we calculated distances among different scores based on correlation coefficients, and constructed a dendrogram of the scores by average linking cluster analysis. The cluster analysis showed that most scores fall into one of two groups– substitution matrix based group

and frequency based group respectively. We also evaluated the scores' performance in predicting DNA-binding sites and found that frequency based scores generally perform best.

So, the same as for catalityc sites [4], conservation and variation scores for prediction of DNA-binding sites can be classified into mainly two large groups. When using a score to predict DNA-binding sites, frequency based scores that also consider a background distribution are most successful.

This work was supported by Grant of Presidium RAS «Molecular and cellular biology» and government contract P1272 of Education and Science Department of Russian Federation.

1. A.Anashkina et al. (2007) Comprehensive statistical analysis of residues interaction specificity at protein-protein interfaces, *Proteins*, **67(4):**1060-1077.
2. A.Anashkina et al. (2008) Geometrical analysis of protein-DNA interactions on the basis of the Voronoi-Delaune tessellation, *Biofizika*, **53(3):**402-406 (in Russian).
3. W.S.J. Valdar (2004) Scoring residue conservation, *Proteins*, **48(2):**227–241.
4. F.Johansson and H.Toh (2010) *BMC Bioinformatics*, **11:**388-398.

Detection of Structurally Invariant Sites in the HIV-1 Third Variable (V3) Loop by Computer-Aided Approaches

Alexander Andrianov[1], Ivan Anishchenko[2], Alexander Tuzikov[2]

[1]*Institute of Bioorganic Chemistry, National Academy of Sciences of Belarus, Belarus* andrianov@iboch.bas-net.by
[2]*United Institute of Informatics Problems, National Academy of Sciences of Belarus, Belarus*

The V3 loop on gp120 from HIV-1 is a focus of many research groups involved in anti-AIDS drug development because this region of the protein is the principal target for neutralizing antibodies and determines the preference of the virus for T-lymphocytes or primary macrophages. Although the V3 loop is a promising target for anti-HIV-1 drug design, its high sequence variability is a major complicating factor. Nevertheless, the occurrence of highly conserved residues within the V3 loop allows one to suggest that they may preserve their conformational states in different HIV-1 strains and, therefore, should be promising targets for designing new anti-HIV drugs. In this connection, the issue of whether these conserved amino acids may help to keep the local protein structure and form the structurally rigid segments of V3 exhibiting the HIV-1 vulnerable spots is very relevant. One of the plausible ways to answer this question consists of examining the V3 structures for their consensus sequences corresponding to the HIV-1 group M subtypes responsible for the AIDS pandemic followed by disclosing the patterns in the 3D arrangement of the variable V3 loops. Because of the deficiency of

experimental data on the V3 structures, these studies may be performed by homology modeling using the high-resolution X-ray and NMR-based V3 models as the templates.

In this work, the 3D structural models for the consensus amino-acid sequences of the V3 loops from the HIV-1 subtypes A, B, C, and D were generated by bioinformatics tools to reveal common structural motifs in this functionally important portion of the gp120 envelope protein. To this effect, the most preferable 3D structures of V3 were computed by homology modeling and simulated annealing methods and compared with each other, as well as with those determined previously by X-ray diffraction and NMR spectroscopy. Besides, the simulated V3 structures were also exposed to molecular dynamics computations, the findings of which were analyzed in conjunction with the data on the conserved elements of V3 that were obtained by collation of its static models.

As a matter of record, despite the high sequence mutability of the V3 loop, its segments 3-7, 15-20 and 28-32 were shown to form the structurally invariant sites, which include amino acids critical for cell tropism. Moreover, the biologically meaningful residues of the identified conserved stretches were also shown to reside in □-turns of the V3 polypeptide chain. In this connection, these structural motifs were suggested to be used by the virus as docking sites for specific and efficacious interactions with receptors of macrophages and T-lymphocytes. Therefore, the structurally invariant V3 sites found here represent potential HIV-1 weak points most suitable for therapeutic intervention.

In the light of the findings obtained, the strategy for anti-HIV-1 drug discovery aimed at the identification of co-receptor antagonists that are able to efficiently mask the structural motifs of the V3 loop, which are conserved in different virus subtypes, is highly challenging. To overcome this problem, an integrated computational approach involving theoretical procedures, such as homology modeling, molecular docking, molecular dynamics, QSAR modeling and free energy calculations, should be of great assistance in the design of novel, potent and broad antiviral agents.

Acknowledgment

This study was supported by grant from the Union State of Russia and Belarus (scientific program SKIF-GRID; № 4U-S/07-111).

An evolutionary space for microbial evolution and community structure analysis

E.V. Pershina[1]; A.S. Dolnik[2]; G. Tamazyan[2]; E.V. Ikonnikova[2];
K.V. Vyatkina[2]; A.G.Pinaev[1]; E.E. Andronov[1]

[1] *ARRIAM, Podbelsky sh., 3, St.-Petersburg, Pushkin 196608, eeandr@gmail.com*
[2] *The Information Management Research Group at the University of St.Petersburg, alexanderdolnik@acm.org*

The modern microbiology has entered the era of metagenomics when the old way to manage sequence data as a list of records is due to be more "integral", allowing investigation of specific and global microbial communities as a whole. A concept of "taxonomic" or "evolutionary" space (ES) where unique nucleotide sequences (e.g. 16S rRNA gene) are presented by dots and the distance between every two dots corresponds to evolutionary distances between these sequences exists as an appealing idea not only for microbial community data analysis, but also for adequate understanding of real evolutionary patterns which may be beyond of reach of traditional approaches (1-4). Taking into consideration the size of the current 16S rRNA gene database (more that 10^6 records) and the number of dimensions required for the corresponding true ES (probably more than 10^5), we understand that the shortest way to construct a "working" ES is to find a low dimensional space approximation preserving main topological features and evolutionary metrics of the true ES.

The aim of this study is to design a working ES with acceptable dimension and evolutionary reasonable metric. Two version of ES were designed by using current bacterial 16S rRNA databases and common statistics. Three experimental sets of the 16S rRNA gene fragment sequences were obtained from the same soil DNA sample with three different primer pairs. The first version of a 100-dimensional ES ("illegal") was constructed for "primer bias" effect measurement: three libraries of 16S rRNA sequences were mapped in ES_1. Despite of "illegal" character of ES_1 we found a good correlation (0,87) between "calculated" (by using ES_1 metric) and "true" distance matrices. The analysis of relation between correlation coefficients and the number of dimensions suggested that satisfactory results can be achieved by using 40 dimensions. Coordinates of the central point were calculated for each sequence set. A primer bias effect with clear taxonomic sense was quantified. A number of integral characteristics of microbial communities and their dynamic are proposed (density, volume, shape, trace, speed etc).

The second version of ES is under construction. It is more rigorous and has been designed for evolutionary studies of 16S rRNA databases. To some extent it has a "cosmological" character. If we take the existence of the last common ancestor for bacteria as a "null hypothese" and make some very rough simplifications, the representation of the global evolutionary process in Bacteria domain in the true ES (perhaps bounded by functional constraints) should look like the Big Bang with one big difference: contrary to the Big Bang the global bacterial community represented in the true ES should have bigger density in peripheral than in the central areas (maybe like a multidimensional diffuse hollow sphere). Strictly speaking our working hypothesis predicts that the global microbial diversity can be presented in the true ES as an irregular fractal with a multidimensional hollow sphere as the basic element. All the data presented are highly disputable but we believe that there is need for more integral ways to presentation and analyses of metagenomic data and the ES concept is one of such opportunities.

The work was supported by Ministry Education and Science of Russian Federation (c. 16.512.11.2132) and Program of fundamental research in SPbSU.

1. Y. Kitazoe, Y. Kurihara, Y. Narita et al. (2001). A New Theory of Phylogeny Inference Through Construction of Multidimensional Vector Space. Mol. Biol. Evol. **18(5):**812–828
2. G.M. Garrity and T.G. Lilburn (2002). Mapping taxonomic space: an overview of the road map to the second edition of Bergey's Manual of Systematic Bacteriology, WFCC Newsletter, **35:** 5-15
3. D. M. Hillis ,T. A. Heath and K. St.John (2005). Analysis and Visualization of Tree Space. Syst. Biol. **54(3):**471–482
4. K. Rudi, M. Zimonja and T. Næs (2006). Alignment-independent bilinear multivariate modelling (AIBIMM) for global analyses of 16S rRNA gene phylogeny. Int. J. Syst. Evol. Microbiol., **56:** 1565–1575

Development of Novel Anti-HIV-1 Agents Based on Glycosphingolipids by Computer Modeling and Chemical Synthesis: β-Galactosylceramide and the Envelope GP120 V3 Loop

Ivan Anishchenko[1], Alexander Andrianov[2], Mikhail Kisel[2], Vasiliy Nikolaevich[2], Vladimir Eremin[3], Alexander Tuzikov[4]

[1]*United Institute of Bioinformatics Problems, National Academy of Sciences of Belarus, Belarus,* anishchenko.ivan@gmail.com
[2]*Institute of Bioorganic Chemistry, National Academy of Sciences of Belarus, Belarus,*
[3]*The Republican Research and Practical Center for Epidemiology and Microbiology, Belarus*
[4]*United Institute of Informatics Problems, National Academy of Sciences of Belarus, Belarus*

The HIV-1 V3 loop plays a central role in the biology of the HIV-1 envelope glycoprotein gp120 as a principal target for neutralizing antibodies, and as a major determinant in the switch from the non-syncytium-inducing to the syncytium-inducing form of HIV-1 that is associated with accelerated disease progression. HIV-1 cell entry is mediated by the sequential interactions of gp120 with the receptor CD4 and a co-receptor, usually CCR5 or CXCR4, depending on the individual virion. The V3 loop is critically involved in this process. Because of the exceptional role of the V3 loop in the viral neutralization and cell tropism, one of the actual problems is that of identifying chemical compounds able to block this functionally crucial site of gp120. According to empirical observations, glycolipid beta-galactosylceramide (beta-GalCer) forming on the surface of some susceptible host cells the primary receptor for HIV-1 alternative to CD4 exhibits a strong attraction to the V3 loop and, for this reason, may be involved in anti-HIV-1 drug studies. In the light of these observations, the use of bioinformatics tools for imitating the process of making the V3/glycolipid complexes may provide a structural rationale for the design of efficient blockers of the functionally important V3 sites.

The objects of this study were to generate the 3D structure model for the complex of V3 with beta-GalCer and, based on the calculation data, to design its water soluble analogs that could efficiently mask the HIV-1 V3 loop followed by their synthesis and medical trials. To this effect, the following problems were solved: (i) 3D structures for the consensus amino acid sequences of the HIV-1 subtypes A and B V3 loops were computed by homology modeling and simulated annealing; (ii) spatial structures of beta-GalCer, as well as of a series of its modified forms were determined by quantum chemistry and molecular dynamics simulations; (iii) supramolecular ensembles of these glycolipids with V3 were built by molecular docking methodology and energy characteristics describing their stability were estimated by molecular dynamics computations; (iv) synthesis of beta-GalCer derivatives that, according to the designed

data, give rise to the stable complexes with V3 was performed, and (v) testing of these compounds for antiviral activity was carried out. From the structural data obtained, the Phe and Arg/Gln amino acids of the gp120 immunogenic crest were revealed to play a key role in forming the complexes of glycolipids with V3 by specific interactions with the galactose residue and sphingosine base respectively. And at the same time, the sugar hydroxyl groups form the H-bonds with the nearby polar atoms of the V3 backbone. Two water soluble analogs of beta-GalCer were also found to display a high affinity to V3 close to that of the native glycolipid. This inference results from the values of binding free energy evaluated for the calculated structures and coincides with the experimental data on the complexes of gp120 with beta-GalCer. The above theoretical findings are in keeping with those of medical trials of the synthesized molecules, which testify to their anti-HIV-1 activity against the virus subtypes A and B isolates.

As a matter of record, the molecules constructed here are supposed to present the promising basic structures for the rational design of novel potent HIV-1 entry inhibitors that could neutralize the majority of circulating indigenous strains.

Computer-Assisted Anti-AIDS Drug Development: Cyclophilin B Against the HIV-1 Subtype A V3 Loop

Ivan Anishchenko[1], Yurij Kornoushenko[2], Alexander Andrianov[2]

[1] *United Institute of Informatics Problems, National Academy of Sciences of Belarus, Minsk, Belarus,* anishchenko.ivan@gmail.com

[2] *Institute of Bioorganic Chemistry, National Academy of Sciences of Belarus, Minsk, Belarus*

The objects of this study originated from the experimental observations, whereby the HIV-1 gp120 V3 loop is a high-affinity ligand for immunophilins, and consisted in generating the structural complex of cyclophilin (Cyc) B belonging to immunophilins family with the virus subtype A V3 loop (SA-V3 loop), as well as in specifying the Cyc B segment forming the binding site for V3, synthetic copy of which may present a promising basic structure for anti-AIDS drug development.

To reach the objects of view, molecular docking of the HIV-1 SA-V3 loop structure determined previously with the X-ray conformation of Cyc B was put into practice by Hex 4.5 program (http://www.loria.fr/~ritchied/hex/) and the immunophilin stretch responsible for binding to V3 (Cyc B peptide) was identified followed by examination of its 3D structure and

dynamic behavior in the unbound status. To design the Cyc B peptide, the X-ray conformation for the identical site of the native protein was involved in the calculations as a starting model to find its best energy structural variant. The search for this most preferable structure was carried out by consecutive use of the molecular mechanics and simulated annealing methods. The molecular dynamics computations were implemented for the Cyc B peptide by the GROMACS computer package (http://www.gromacs.org/).

As a result, the supramolecular structure of Cyc B with V3 was built by computer modeling tools and the immunophilin-derived peptide able to effectively mask the structurally invariant V3 segments, which include the functionally crucial amino acids of the HIV-1 gp120 envelope protein, was constructed and analyzed (Figure).

Starting from the joint analysis of the results derived with those of the literature, the generated peptide was suggested to offer a promising basic structure for making a reality of the protein engineering projects aimed at developing the anti-AIDS drugs able to stop the HIV's spread.

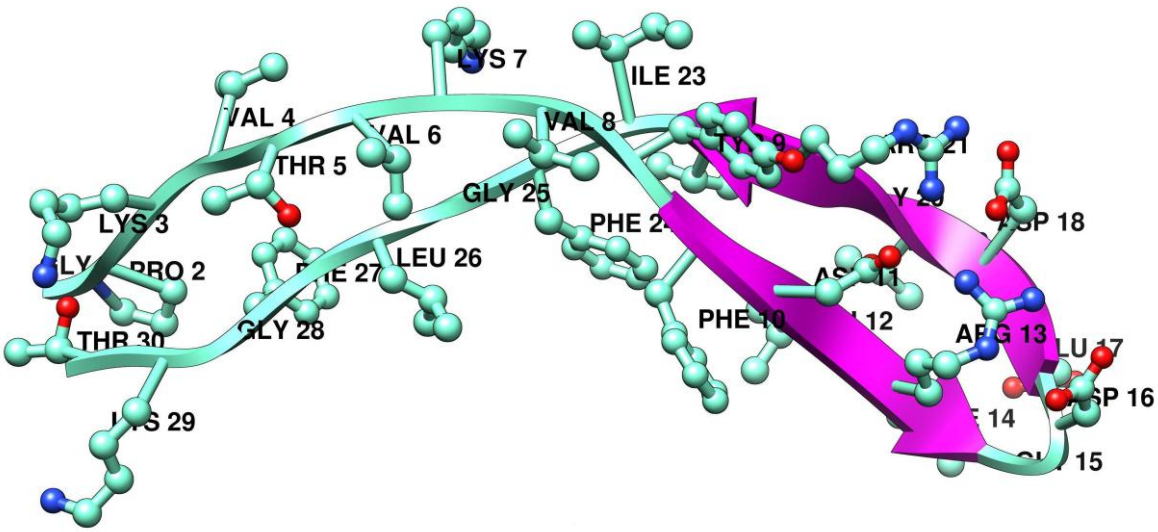


Figure. 3D structure of the CycB peptide generated by bioinformatics tools

Acknowledgment

This study was supported by grant from the Union State of Russia and Belarus (scientific program SKIF-GRID; № 4U-S/07-111), as well as from the Belarusian Foundation for Fundamental Research (project X10-017).

TEpredict – software for predicting T-cell epitopes: an update

Denis Antonets

State Research Center of Virology and Biotechnology "Vector", Russian Federation, antonec@yandex.ru

It is well known that CD8+ T-cell epitopes play crucial role in antiviral and anticancer immunity. Reliable prediction of CTL epitopes remains one the most important goals of immunoinformatics as accurate in silico identification of potent T-cell epitopes could drastically reduce materials and time consumption compared to the traditional experimental approaches of epitope discovery. The main aim of this work was the development of new statistical models for predicting peptide binding to different allomorphs of MHC class I molecules to update TEpredict software [1].

New models for predicting affinity of peptide-MHC binding were constructed by means of either partial least squares (PLS) regression [2] or with recently developed sparse partial least squares (SPLS) technique, that was shown to outperform convenient PLS regression method [3]. Since the majority of known CTL epitopes have nine aminoacids in length, only nonapeptide:MHC binding data, collected from Immune Epitope Database (IEDB; http://www.immuneepitope.org), was used to build predictive models for 35 allelic variants of HLA class I molecules.

The recently developed amino acid similarity matrix PMBEC, derived from experimentally determined peptide:MHC binding data [4], was used to parameterize peptides. Encoding of amino acid residues with rows of PMBEC matrix was found to provide better predictive models as compared to sparse-encoding and BLOSUM62-based parameterization [4]. Since results of principal component analysis and factor analysis of PMBEC matrix suggested that amino acids could be described well with 5-7 coordinates, the original 20-dimensional amino acid parameterization space was transformed into the set of 5-15 dimensional spaces by means of mutual information-based independent component analysis (PearsonICA package for R) [5] or using either isometric multidimensional scaling [6] or Sammon's method [7] from the R package MASS. All produced scales were applied for parameterizing nonameric peptides and the performance of corresponding PLS models was comparatively assessed.

As it was expected, the models built with PMBEC parameterization were shown to outperform those built with sparsely-encoded peptides, but although they performed well on testing data, these models were much more complex (as they have 180 predictor variables and the majority of such PLS models had more than 14 hidden components) and they had less predictive power in cases with small training datasets as compared to the models that were built using ICA-transformed amino acid parameterization scales. All PLS models that were built with ICA-produced scales with dimensionality > 10 were shown to slightly outperform sparse-

encoding-based models and PMBEC-based ones both in terms of the area under the ROC curve (AUC) and Pearson's correlation coefficient. In general, 11D parameterization scheme produced with ICA was almost as efficient as PMBEC and for some HLA alleles (HLA-A*2402, HLA-A*2403) resulted in more accurate predictive models. It's of interest that, as judged by AUC, for some HLA alleles the most qualitatively accurate predictive models were produced using 7D parameterization scales. AUC values, obtained by testing new models exploiting 11D parameterization scheme, ranged from 0.77 to 0.98 with mean and median values equal to 0.91.

Results of comparative testing of produced models and updated TEpredict software could be found at http://tepredict.sourceforge.net/update.html.

1. D.V. Antonets, A.Z. Maksyutov (2010) TEpredict: software for T-cell epitope prediction, Mol Biol (Mosk), 44:130–139.
2. B.-H. Mevik, R. Wehrens (2007) The pls Package: Principal Component and Partial Least Squares Regression in R, Journal of Statistical Software,18:1–24.
3. H. Chun, S. Keles (2010) Sparse partial least squares for simultaneous dimension reduction and variable selection, Journal of the Royal Statistical Society - Series B, 72:3–25.
4. Y. Kim et al. (2009) Derivation of an amino acid similarity matrix for peptide: MHC binding and its application as a Bayesian prior, BMC Bioinformatics, 10:394.
5. J. Karvanen, V. Koivunen (2002) Blind separation methods based on Pearson system and its extensions, Signal Processing, 82:663–673.
6. T.F. Cox, M.A.A. Cox (1994, 2001) Multidimensional Scaling. Chapman & Hall
7. J.W. Sammon (1969) A non-linear mapping for data structure analysis, IEEE Trans. Comput., C-18:401–409.

Theoretical study of structural features of variola virus CrmB protein

Denis Antonets, Tatyana Nepomnyashchikh, Sergei Shchelkunov

[1]*State Research Center of Virology and Biotechnology "Vector", Russian Federation, antonec@yandex.ru*

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

Models of TNF receptor domains of VARV- and CPXV-CrmB were constructed using Modeller (9v2) software (http://salilab.org/modeller) and validated with ProCheck [1]. Models of ligand-receptor complexes of VARV and cowpox virus (CPXV) CrmBs with hTNF (1TNF) and mTNF (2TNF) were produced by superimposing corresponding molecules onto the crystall structure of human TNF receptor I (p55) complex with lymphotoxin (1TNR). All constructed models were then energy minimized using either NOC (http://noch.sourceforge.net) or FoldX (http://foldx.crg.es). Stability of ligand-receptor complexes was predicted either with FoldX or using residue-level pairwise potentials BETM990101 [2]. Analysis of these models with either

FoldX or with BETM990101 pair potentials revealed that mTNF should bind to CPXV-CrmB with higher affinity than hTNF. VARV-CrmB was predicted to bind both cytokines with higher affinity than CPXV-CrmB; CPXV-CrmB was predicted to bind hTNF(R31Q) with significantly higher affinity than wild type hTNF. Using FoldX both CrmBs were predicted to less efficiently bind to hTNF(E127Q), than to the wild type hTNF. All these findings were then approved by experimental evaluation of VARV- and CPXV-CrmB proteins ability to inhibit cytotoxic action of mTNF, hTNF, hTNF(R31Q) and hTNF(E127Q) on L929 murine fibroblast cells. Predicted stability of modelled ligand-receptor complexes of both CrmBs with selected TNFs was found to be in good qualitative agreement with experimental data. Produced models will be used for designing mutant forms of VARV-CrmB with higher affinity towards hTNF.

Recently VARV CrmB protein was also shown to bind with high affinity several chemokines which recruit B- and T-lymphocytes and dendritic cells to sites of viral entry and replication and ability to bind chemokines was shown to be associated with unique C-terminal domain of CrmB protein [3]. This domain named SECRET (Smallpox virus-Encoded Chemokine Receptor) is unrelated to the host proteins and lacks significant homology with other known viral chemokine-binding proteins or any other known protein. Using I-TASSER server [4], which was shown to be the best server for predicting spatial structures of proteins according to results of CASP7 (CASP - Critical Assessment of protein Structure Prediction) and CASP8 competitions, we obtained the model of VARV-CrmB SECRET domain. As the best template for modelling SECRET spatial structure I-TASSER has chosen the structure of CPXV vCCI protein (PDB ID: 1CQ3) belonging to the family of poxviral type II chemokine-binding proteins (vCkBPsII) which are encoded by almost all known members of Orthopoxvirus and Leporipoxvirus genera. As other members of vCkBPsII SECRET was predicted to be beta-sandwich, composed by two parallel beta-sheets connected by several loops and its ligand-binding surface was predicted to have prominent electronegative charge required for binding to positively charged conservative amino-acid residues of chemokines. Thus our results suggest that SECRET should be included into the family of poxviral type II chemokine-binding proteins and that it might have been evolved from the vCCI-like predecessor protein.

This work was supported by Russian Foundation for Basic Research (grant #09-04-00055a) and by the government of Novosibirsk region.

1. R.A. Laskowski et al. (1993) PROCHECK: a program to check the stereochemical quality of protein structures. J. Appl. Cryst., 26:283–291.
2. M.R. Betancourt, D. Thirumalai (1999). Pair potentials for protein folding: choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. Protein Sci. 8:361–369.
3. R.A. Alejo et al. (2006) A chemokine-binding domain in the tumor necrosis factor receptor from variola (smallpox) virus. Proc Natl Acad Sci USA, 103:5995–6000.
4. Y. Zhang (2008) I-TASSER server for protein 3D structure prediction. BMC Bioinformatics, 9:40.

Finding and sorting out frameshifts in 1100+ prokaryotic genomes: programmed frameshifts, pseudogenes and sequencing errors

Ivan Antonov[1], Mark Borodovsky[2]

[1]*Division of Computational Science and Engineering, Georgia Institute of Technology, 801 Atlantic Drive, Atlanta, GA, USA 30332, ivan.antonov@gatech.edu*
[2]*Department of Biomedical Engineering and School of Computational Science and Engineering, Georgia Institute of Technology, 313 Ferst Drive, Atlanta, GA, USA 30332, borodovsky@gatech.edu*

An *ab initio* frameshift prediction program GeneTack [1] was applied to 1,106 prokaryotic genome sequences. Overall, 206,991 frameshifts have been predicted. All genes containing predicted frameshifts (fs-genes) were conceptually translated into proteins (fs-proteins).

The fs-proteins were further used as queries for the BLASTp search against NCBI nr database to detect homologous fs-proteins in other species; we also attempted to locate Pfam domains in fs-proteins. Notably, for 10,961 fs-proteins both BLASTp and Pfam searches identified homologous proteins combining translations of both sides of the broken frame, thus the corresponding frameshift predictions were validated .

Next, a database of all fs-proteins was built and "all-against-all" BLASTp search was done in order to clump together orthologous fs-proteins; as a result, the fs-proteins were grouped into 19,666 clusters of orthologous fs- proteins (COFs).

Since the GeneTack performance, as assessed earlier, delivers 85.8% sensitivity and 68.2% specificity in frameshift detection, we expect that almost 1/3 of the predictions are false positives.

Further down the road our goals were i/ to filter out false positive predictions and ii/ to determine a nature of true positive predictions, i.e. to classify them as ones caused by either random sequencing error or representing a true sequence feature which could inactivate a gene (pseudogene) or be involved in gene regulation (programmed frameshift).

The fs-proteins produced by recovering a sequencing error do not have homologs among other fs-proteins. Therefore, 100,991 of one element COFs (singletons) are likely to represent sequence errors 5,523 of them validated by both BLASTp and Pfam.

Frameshifts that show conservation in homologous genes are indicative of a conserved frameshift mutation in the lineage. Still, false positive frameshift predictions related to adjacent/overlapping gene pairs with conserved co-location could also form clusters.

Presence of a specific conserved motif situated close to the frameshift site is an important feature of clusters of programmed frameshifts discriminating them from clusters of fs-proteins derived from pseudogenes and from clusters of false positive fs-proteins.

Not only a motif itself but also its phasing with respect to the reading frame is crucial for proper functioning of a programmed ribosomal frameshift (PRF). The phasing was taken into account in a new algorithm that identifies possible programmed frameshifts motifs better than standard motif searching algorithm, e.g. MEME.

To identify clusters of fs-proteins which expression is regulated by programmed frameshifts we considered multiple features such as cluster size, motif strength and type, average Ka/Ks ratio (indicative of pseudogenes), etc. Technically, each cluster makes a point in N-dimensional space where N is number of different features. A clustering algorithm was applied to separate clusters of fs-proteins with different origin, particularly ones that require programmed frameshifts for their expression regulation. In the results section we describe our findings of programmed frameshifts and pseudogenes and their distributions among prokaryotic species.

We would like to thank Pavel Baranov for very useful discussions of biological features of programmed frameshifts.

1. Antonov I., M. Borodovsky (2010) GeneTack: frameshift identification in protein-coding sequences by the Viterbi algorithm, *J Bioinform Comput Biol.*, **8(3):** 535-51.

GeneTack website: http://topaz.gatech.edu/GeneTack/

3D-CSD: a resource of 3D structure of catalytic sites and prediction of catalytic sites in proteins by approximate sub-graph isomorphism

Seyed Shahriar Arab[1], Mohammad Ebrahim Abbasi Dezfouli[2], Najmeh Hosseynimanesh[2]

[1] *School of Computer Science, Institute for Research in Fundamental Science (IPM), Tehran, Iran, shahriar@ipm.ir*
[2] *Sharif University of Technology, Electrical Engineering Department, BiSIPL (Biomedical Image and Signal Processing Lab), Tehran, Iran*

The prediction of catalytic residues is a key step in understanding the function of enzymes and classifying them. This is a very challenging job, because an Amino Acid can appear in a variety of active sites. The biological activity of a protein usually depends on the presence of a small number of functional residues. Identifying these residues from the sequence has many applications. Classically, strictly conserved residues are predicted to be functional, but often conservation patterns are more complicated. There are a lot of algorithms to predict functional site, but few are available via publicly accessible application. Here, we represent 3D-CSD, a database containing 3D structure of active sites, and a tool for predicting the active site based on spatial shape. and the type of functional residues.

3D-CSD is a database consists of information about the enzymes, related function and the specifications of the active site in enzymes.

Proteins and active sites are represented by a connected graph G(V,E) in which the nodes (V) represent the Cα atoms of each residue and the edges (E) of the graph are weighted with the inverse of the distance between each two residues.

We use an approximate sub-graph isomorphism algorithm for finding the active site in the query protein. This algorithm searches the graph model of the query protein to find the active site sub-graph.

This data can be used to predict the function and the active site of a query protein. This capability of the program can increase our knowledge about the enzymes. Also it can be used for the hypothetical proteins that their structure is known with no information about their function. The structure of these proteins has been determined by X-Ray, NMR or even prediction methods. This software is useful to infer the active site and function of enzymes, and also it has many applications in designing new enzymes.

Acknowledgements: This research was in part supported by a grant from IPM (no. CS1385-102).

1. P. Aloy, E. Querol, F. X. Aviles, and M. J. Sternberg (2001) Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking, *J Mol Biol*, **311**: 395-408 .
2. J. R. Bradford and D. R. Westhead (2005) Improved prediction of protein-protein binding sites using a support vector machines approach, *Bioinformatics* **21**: 1487-1494 .
3. T. Bray, P. Chan, S. Bougouffa, R. Greaves, A. J. Doig, and J. Warwicker (2009)SitesIdentify: a protein functional site prediction tool, *BMC Bioinformatics*, **10**: 379.
4. G. Cheng, B. Qian, R. Samudrala, and D. Baker (2005) Improvement in protein functional site prediction by distinguishing structural and functional constraints on protein family evolution using computational design, *Nucleic Acids Res.*, **33**: 5861-5867.
5. A. H. Elcock (2001) Prediction of functionally important residues based solely on the computed energetics of protein structure, *J Mol Biol*, **312**, 885-896.
6. R. A. Laskowski, N. M. Luscombe, M. B. Swindells, and J. M. Thornton (1996) Protein clefts in molecular recognition and function, *Protein Sci*, **5**: 2438-2452.
7. M. Ota, K. Kinoshita, and K. Nishikawa (2003) Prediction of catalytic residues in enzymes based on known tertiary structure, stability profile, and sequence conservation, *J Mol Biol*, **327**: 1053-1064.
8. A. R. Panchenko, F. Kondrashov, and S. Bryant (2004) Prediction of functional sites by analysis of sequence and structure conservation, *Protein Sci* **13**: 884-892.
9. C. T. Porter, G. J. Bartlett, and J. M. Thornton (2004) "The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data, *Nucleic Acids Res.*, **32**: D129-D133.
10. S. Sankararaman, F. Sha, J. F. Kirsch, M. I. Jordan, and K. Sjolander (2010) Active site prediction using evolutionary and structural information, Bioinformatics **26**: 617-624.

Kinetic model explains correlation between DNA methylation and tissue-specific alternative splicing

Artem Artemov[1], Dmitri Pervouchine[1], Alexander Favorov[2], Andrey Mironov[1]

[1]*Department of Bioengineering and Bioinformatics, Moscow State University, Vorobiovy Gory 1-73, Moscow 119992, GSP-2 Russia, artemov@bioinf.fbb.msu.ru*
[2]*GosNIIGentika, Moscow*

There is increasing evidence that alternative splicing provides significant contribution to the diversity of tissue-specific gene products. A number of recently emerged experimental results support the role of epigenetic modifications and chromatin structure in the regulation of alternative splicing. Assuming that at least some of the splicing events occur co-trancriptionally, the regulation can be explained by so called kinetic model. The kinetic model suggests that, when spliceosome binds the donor splice site during transcription, the chromatin state can affect the choice of the acceptor splice site by decreasing the RNA polymerase elongation rate, thereby providing competitive advantage to the 5'-most acceptor splice site compared to the other splice sites located further downstream.

In the current work we performed a large-scale analysis to find associations between tissue-specific patterns of alternative splicing and tissue-specific patterns of DNA methylation that would be consistent with the kinetic model. For simplicity we chose the particular case of

cassette exons. We observed that DNA regions corresponding to introns which follow cassette exons tend to have higher average methylation scores compared to introns located downstream of constitutive exons. We have shown that this effect is not due to the difference in downstream intron lengths, which might as well affect the lag time between splice site synthesis events. We have also shown that higher average methylation downstream of cassette exons remains statistically significant when compared to the average methylation of the upstream region. The elevated DNA methylation downstream of constitutive exons doesn't seem to be localized in any particular region between alternative acceptor sites but is rather uniformly distributed along the downstream intron. We also find statistical evidence for correlation between cassette exon's inclusion rate and methylation of its downstream intron; however, the statistical significance of these findings is rather weak due to lack of the data. Genes that contain a cassette exon followed by a highly methylation intron tend to be classified as being involved in regulation of biological process (GO:0050789) and developmental processes (GO:0032502).

Taken together, our results support the kinetic model for regulation of alternative splicing by changes in chromatin structure resulting from tissue-specific DNA methylation.

Multiple structural alignment of α/β hydrolase-fold enzymes and bioinformatic analysis of catalytically important residues

Vladimir Arzhanik[1], Eugeny Kirilin[1], Dmitry Suplatov[2], Vytas Svedas[1,2]

[1]*Faculty of Bioengineering and Bioinformatics,MSU, Russian Federation, arzhanik_work@mail.ru*
[2]*Belozersky Institute of Physicochemical Biology, MSU, Russian Federation*
* Corresponding author - *vytas@belozersky.msu.ru*

Comparative bioinformatics is the cornerstone of computational approaches to understanding structure-functional relationship in enzymes. The members of α/β hydrolase-fold superfamily represent a functionally diverse group of enzymes with common structural organization that appear to have lost sequence similarity during natural selection and specialization from a common ancestor. At the same time their active site structures in general remain conserved while other parts may largely differ. It is therefore expected that three-dimensional alignment will provide more significant clues to protein function, properties and evolution than sequence alignment alone. Here we report the largest so far multiple structural comparison of α/β hydrolase-fold superfamily enzymes and analysis of catalytically important residues considering distribution of amino acid types among enzymes with different catalytic properties.

A computer algorithm has been developed for high-throughput structural comparison of homologous enzymes. Comparative analysis of the most functionally significant parts of enzyme structures – the active sites – is suggested as a source of new understanding of structure-functional relationship in α/β hydrolases. On the first step a library of active site structures of α/β hydrolases – amino acids involved in catalysis together with residues forming the active site cavity and thus potentially involved in mechanical aspects of enzyme behavior by interacting with the substrate or catalytic machinery – was created using previously reported procedure [1]. Then a representative set of structures as the basis for comparison of distinct subfamilies was selected using 95% sequence similarity threshold. On the next step a superimposition matrix was created from pairwise comparisons of representative structures. Finally, amino acid positions conserved between the structures were determined and clustered to form the common core alignment.

Multiple alignment of α/β-hydrolase superfamily was created on the basis of 238 non-redundant structure set of enzymes with lipase, esterase, hydroxynitrile lyase, epoxide hydrolase, peptidase, dienelactone hydrolase and dehalogenase activities. To the best of our knowledge this is the largest structural alignment of the superfamily reported so far. Comparative analysis

revealed a major structural similarity of active site regions while the most significant fit was observed near the catalytic triad residues. While catalytic His was found to be conserved among all α/β-hydrolase enzymes, the nucleophile and catalytic acid were identified as subfamily-specific positions - residues with a tendency to be conserved only within subfamilies of enzymes, but different between subfamilies. Nucleophile position can be occupied by Ser, Asp or Cys. Ser was found to be common for the majority of activities explored. Asp is considered as stronger nucleophile compared to Ser and contributes to S_N2 reaction mechanism of epoxide hydrolases and dehalogenases. Dienelactone hydrolases have Cys as a nucleophile due to substrate-assisted catalysis, where a functional group in the substrate controls the protonation state of the nucleophile. Catalytic acid was found to be represented by Asp or Glu. Asp was shown to be the most common in this position, while Glu was found in acetylcholine esterases and some carboxypeptidases. Role of this substitution yet remains unknown. It was shown, that the origin of catalytic acids is not always homologous – in the majority of structures it is hosted by β7 sheet, though in 15% of cases it can be found on β6 sheet of the α/β-fold. Proteins with α/β-hydrolase fold that do not have reported enzymatic activity like gibberellin receptors, cell adhesion proteins, signaling proteins etc. were also considered. It was shown that catalytic triad positions are occupied by residues that do not support catalytic activity, like Gly in nucleophile position or Val in place of a catalytic histidine.

This work was supported by Russian Ministry of Science and Innovation (contract № 02.740.11.0866)

1. D.A.Suplatov, V.K.Arzhanik, V.K. Švedas; Acta Naturae, 2011, 3(1), 99-105.

HFE haplotypes: enigmas of multiple isoforms.

Vladimir Babenko, Svetlana Mikhailova, Aida Romaschenko
Institute of Cytology and Genetics, Russian Federation, bob@bionet.nsc.ru

HFE (OMIM: 613609, 612635) is required for normal regulation of hepcidin synthesis in liver and hepcidin-mediated iron export from macrophages, enterocytes, and hepatocytes (Makui et al., 2005).

In our previous study (Mikhailova et al., 2010) we concentrated on snp haplotypes referring to the second haploblock from the two total encompassing the gene.We analysed the distribution of HFE haplotypes across various indigenious as well as contemporary populations of Russia. The dramatic variance of haplotype frequences was found, featuring distinct Asian and European haplotypes pattern that is confimed by HapMap data (www.hapmap.org). On the other hand, the multidude of HFE isoforms (13 Refseq entries) are observed and the range of them confirmed experimentally (Martins et al., 2011). Analysis of haplotype pattern points at their putative linkage to the isoforms frequencies implying their splicing-specific effect.

Current work is devoted in further research of HFE expression specifics by means of auxilliary data sources namely:

a) Affymmetrix exon chip data (http://www.affymetrix.com/products_services/research_solutions/methods/wt_geneexpression_altsplicing.affx). We found out that exons 2 and 4 are the most variable ones considering expression across 11 tissues . Another point was that long as well as short isoforms are expressed depending on the tissue. It complies with current experimental evidence (Martins et al., 2011).

b) ENCODE histone modification University of Washington and Broad Institute data (www.encode.org).

We checked the chromatin state profile within the gene in 8 tissues and observed distinct chromatin state pattern which could accommodate in alternative splicing outcome. In particular the H3K4Me1 histone modification enrichment encompasses haploblock 2 in a tissue – specific manner, that is, it was observed in GM12878 cell line only, peaking in the vicinity of exon 4 upstream region in this cell line possibly implicating termination site markup for the soluble isoform (Martins et al., 2011). A sharp increase has been observed in the downstream region of exon 1 in K562 cell line. H3K4Me3, the promoter-associated histone mark, peaked at the original TSS. H3K36Me3 is observed over the whole gene body in an exon-specific manner.

Overall, we conclude that haplotype blocks in case of HFE correlate with distinct chromatin states, which may afford HFE expression in a tissue-specific manner. In turn, specific chromatin marks exons as was previously reported in (Schwartz, Ast, 2010), stressing in this way the chromatin implication in splicing. Further experimental verification is underway to assess the degree of chromatin implication in HFE expression.

The work is partly supported by RFBR grants (11-04-01206-a) and SB RAS integration grant N115.

1. Makui H, Soares RJ, Jiang W, et al. (2005) Contribution of Hfe expression in macrophages to the regulation of hepatic hepcidin levels and iron loading. Blood 106:2189–2195.
2. Mikhailova SV, Babenko VN, Voevoda MI, Romashchenko AG. The ethnospecific distribution of the HFE haplotypes for IVS2(+4)t/c, IVS4(-44)t/c, and IVS5(-47)g/a in populations of Russia and possible effects of these single-nucleotide polymorphisms in splicing. Genet Test Mol Biomarkers. 2010. 14(4):461-9.
3. Martins R, Silva B, Proença D, Faustino P. Differential HFE gene expression is regulated by alternative splicing in human tissues. PLoS One. 2011;6(3):e17542.
4. Schwartz S, Ast G. Chromatin density and splicing destiny: on the cross-talk between chromatin structure and splicing. EMBO J. 2010. 29(10):1629-36

Non-Coding RNAs: The Cell's Dark Matter

Rolf Backofen

University of Freiburg, Germany, backofen@informatik.uni-freiburg.de

During the last few years, a multitude of regulatory non-coding RNAs (ncRNAs) have been discovered. Many of these act as post-transcriptional regulators by base pairing to a target mRNA, causing mRNA cleavage or translational repression or activation. We will discuss two problem related to ncRNA.

The first task is related to the problem of detecting and classifying ncRNAs. The main basis for this task are comparison methods that are based on both sequence and structure to determine conserved RNA-motifs, and we will dicuss problems and approaches related to this problem.

The second taks is the computational detection of possible targets since experimental verification of targets is difficult. Many existing target prediction programs neglect intra-molecular binding, while other approaches are either specialized to certain types of ncRNAs or too slow for genome-wide searches. We introduce a new fast and general approach to the prediction of RNA-RNA interactions incorporating accessibility of target sites as well as the existence of a user-definable seed. We further discuss approaches that can handle more complex RNA-RNA interaction structures.

Dynamics of amyloidogenic peptide oligomers

Alka Srivastava and Petety V. Balaji

Department of Biosciences and Bioengineering, Indian Institute of Technology Bombay, Powai, Mumbai 400 076, India; balaji@iitb.ac.in

The formation of amyloid fibrils has been associated with a large number of pathological conditions. Study of amyloid fibrils is of interest not only from therapeutical point of view but also from an intellectual point since nearly all the proteins can form amyloids, under some conditions or other. Various aspects of fibril formation and their dissociation into monomers / oligomers is being studied and one of these aspects is the nucleation step. The formation of a 'critical nucleus' is the rate determining step in fibril formation. Addition of pre-formed critical nucleus, or seeding, is known to accelerate fibril formation. The question is what is the structure of the critical nucleus? Experimental methods often are not able to provide atomic level details and dynamics of the critical nucleus since they have a very short half-life. Molecular dynamics simulations have been used extensively to bridge this short-coming. In the present study, the dynamics and stability of pre-formed aggregates of an amyloidogenic peptide has been investigated by molecular dynamics simulations. Several pre-formed aggregates, which differ from each other in their zwitterionic status, size, topology and organization have been simulated. The total duration of the simulations is 3.4 microseconds. The data leads to the inference that the critical nucleus need not be a single species, but instead a heterogeneous mixture of oligomers differing in their size and the way peptides are arranged with respect to each other. It is observed that irrespective of whether the termini are charged or neutral, the cross-beta strands adopt a twist which gives the fibrils the characteristic screw symmetry. Stabilization of the oligomers can be either through the side chain interactions (e.g., by forming double layers) and/or by the interaction of the termini, in addition to the main chain hydrogen bonds. It is also found that the aggregates can dissociate in many different ways.

Gene expression profile of the tumor as a composite biomarker

Ancha Baranova[1], Wang Lei[2], Alessandro Giuliani[3], Ganiraju Manyam[4]

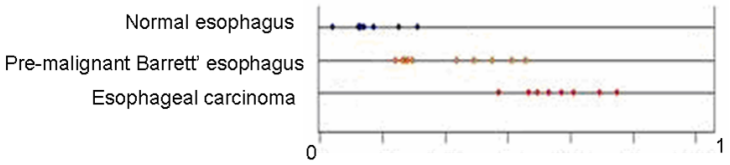
Ancha BARANOVA[1], Wang LEI[2], Alessandro GIULIANI[3], Ganiraju MANYAM[4]

[1]*Russian Center of Medical Genetics RAMS, Moscow, Russia, abaranov@gmu.edu*
[2]*School of Systems Biology, George Mason University, Fairfax, VA, USA*
[3]*Istituto Superiore di Sanità, Viale Regina Elena 299 00161 - Roma*
[4]*The UT MD Anderson Cancer Center, Houston, TX, USA*

To date, the quantification of the diagnostic and prognostic biomarker molecules in the human serum and tissues remains the primary means of enhancing the clinician's ability to predict and detect cancer before it spreads and to predict the outcome of treatment. Importantly, with innumerable molecular markers in development, the discovery of novel standalone markers with acceptable sensitivity and specificity is an extremely rare event. The ideal molecular marker would be one that is inherently related to the process of tumorigenesis or to the defense mechanisms of the individual. However, the same traits of sensitivity and specificity may also allow biomarker molecules to serve as tumor antigens and in such case one would expect that expression of such molecules would be rapidly eliminated in the evolutionary processes within a tumor cell population.

The conventional method to overcome the problem of relatively low sensitivity and specificity of newly discovered biomarkers is to combine them into biomarker panels. However, in many cases these biomarker panels suffer from relatively low reproducibility of results in independently collected sets of samples. This is especially true for the mRNA biomarkers identified by microarray experiments. Here we challenge the biomarker paradigm by developing a distance measure between the entire gene expression profile of a tumor and the center of the space occupied by normal samples. We also introduce a innovative concept of the distance measure between given expression profile and the center of n-dimension space occupied by the set of reference samples with normal (or standard) histology. This novel concept allows one to depart from the classical two-bin prediction model (e.g. "bad prognosis/good prognosis") as it produces a continuous prognosis model, where each sample is located in the neighborhood of other samples analyzed post-hoc and associated with known survival. Here we show that the whole-transcriptome genome based distances calculated using Pearson correlation coefficients provide easy visualization of the relative degree of the malignancy characteristic for studied samples. In all studied datasets, on average, tumors were further away from the Normal Sample Space than the paired samples with normal histology. The distance analysis demonstrated

remarkable behavioral invariance observed in eighteen independent tumor data sets and provided a robust validation of this approach.



Sample chart depicting results of the distance analysis for normal, pre-malignant and malignant samples of esophageal origin

The concept of distance analysis is not limited to cancer as it could be generalized to quantify the departure of any given sample from its reference set, i.e. tissue sample of aged persons from reference of non-aged, samples of insulin resistant tissues from normally functioning tissues, and even model cell lines that drift away from the standard phenotype. If successful, this unconventional approach will shift the tumor biomarker paradigm from expression biomarker panels associated with low reproducibility, to the distance analysis of robust molecular portraits. We also foresee the application of the distance analysis method as a novel technique for the direct evaluation of the departure of a particular batch of model cell line from its original phenotype, thus, replacing the commonly used STR analysis that is sensitive to the genetic, but not to the epigenetic changes in cell lines rapidly acquiring novel chromatin aberrations *in vitro*. The proposed distance analysis is versatile in its application as it will be equally attributable to gene expression profiles collected both by microarrays and by RNA-seq platforms.

Population genetic analysis of ongoing two-nucleotide codon substitutions in D.melanogaster

Mariya Baranova[1], Georgii Bazykin[2], Alexey Kondrashov[3]

[1]*M.V. Lomonosov Moscow State University, Russian Federation, manyashka06@mail.ru*
[2] *Institute for Information Transmission Problems of the Russian Academy of Sciences (Kharkevich Institute), Russian Federation, gbazykin@iitp.ru*
[3]*University of Michigan, United States, kondrash@umich.edu*

Fitness landscape, the function which relates fitness to genotype, can be visualized as isolated peaks separated by valleys. One of the interesting questions in evolutionary biology is whether it is possible for evolving organisms to overpass valleys of low fitness variants to achieve high fitness. For mitochondrial tRNA, it was shown that double substitutions leading to a Watson–Crick switch in complementary pairs at stem region are strongly correlated, and the intermediate variant appears only at low frequencies (1).We study an analogous effect in protein coding sequences. Comparison of codons that differ between two species (2) or two individuals of the same species (3) at two non-synonymous nucleotide sites showed that both substitutions often occur in the same linage. Here, we use the polymorphism of 162 lines of D.melanogaster to analyze the polymorphic codons separated by two nucleotide substitutions. When these two codons encoded the same amino acid (serine), the intermediate variant was observed in 17% of sites, while it was expected at 38% of sites, implying selection against the intermediate variant. For other pairs of nonsynonymous substitutions, the relative fitnesses of the two amino acids are generally not known. Nevertheless, we could assess how the frequency of the intermediate variant depended on the absolute differences in amino acid properties between the three variants. When the two amino acids separated by two nucleotide substitutions were similar, and the intermediate variant was different, the sites with the intermediate variant were rare. Conversely, when these two amino acids were different, we observed the intermediate variant more frequently. Next, we used the codons of D. simulans and D. yakuba to reveal the ancestral state for the two-substitution codons. We found sites where the ancestral and derived amino acids occur simultaneously, and counted the fraction of sites containing the intermediate variant. We found that part of substitutions, which happen without intermediate variant at all, is equal for exons and introns. So, tunneling effect is not specific for exons. But among sites with ancestral and derived amino acids we can see intermediate variant in 12% of exons' sites, compared to 22% of sites in introns. Therefore, in exons, the second substitution follows the first one faster than in introns.

Phylogenetic utility of the low-copy nuclear gene LFY intron 2 in plant molecular phylogenetics as exemplified in Astragalus (Fabaceae)

László Bartha[1], Nicolae Dragoş[1], Attila Molnár V.[2], Gábor Sramkó[2]

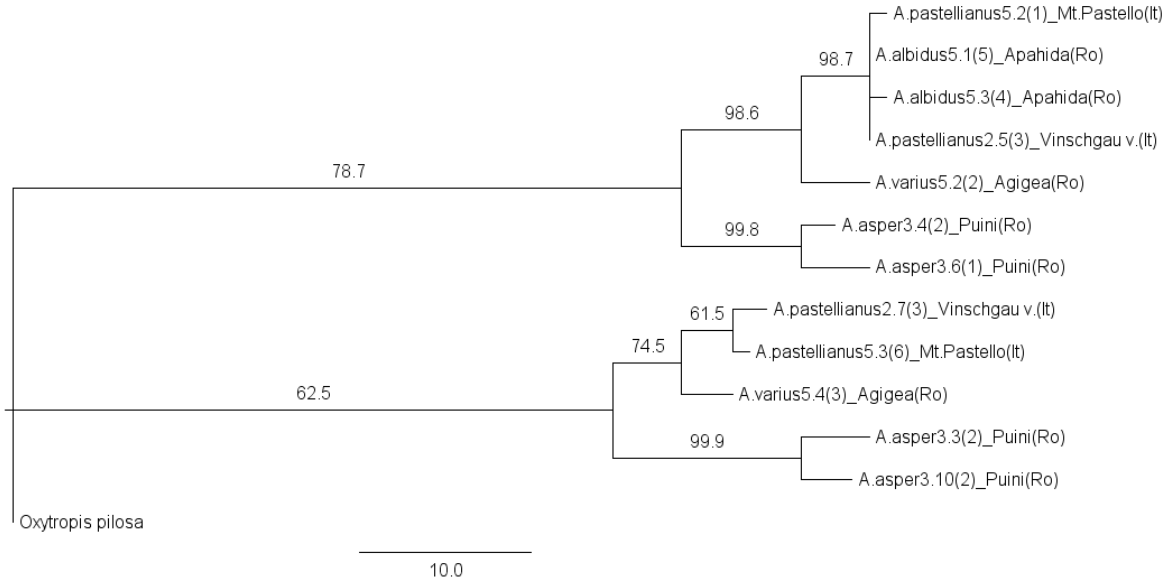
[1] *Babeş-Bolyai University, Str. Republicii, nr. 44., 400015, Cluj-Napoca, Romania, laszlo.bartha@ubbcluj.ro*
[2] *University of Debrecen, Egyetem tér 1., Debrecen, Hungary*

Section Dissitiflori of the diverse genus *Astragalus* has several European members with conservational interest. Nonetheless, the phylogenetic relationships in general are scarcely known in this section. Our recent molecular survey – utilizing the multi-copy plant phylogenetic marker nuclear ribosomal internal transcribed spacer (nrITS) region – found evidence for intra-individual polymorphism that hindered straightforward phylogenetic reconstruction. This was probably due to the polyploid nature of our target species, which can give rise to concerted evolution, intra-individual recombination, and unequal gene-conversion. On the other hand, our results hint at ancient hybridization in the group, which requires usage of nuclear regions. The object of this study was to test the phylogenetic utility of the low-copy nuclear gene LFY 2nd intron for determining phylogenetic relationships in European taxa of the genus *Astragalus* sect. Dissitiflori. Since low-copy genes are thought not to be susceptible to the abovementioned disadvantages of multi-copy genes, it gave promise of an alternative to the more conventional nrITS.

Taxon sampling included two subspecies of *A. vesicarius* s. str. (*A. pastellianus* and *A. albidus*), *A. varius* and *A. asper*, plus *Oxytropis pilosa* as outgroup. Forward and reverse primers were designed using the Fabaceae LFY exon 2 and exon 3 sequences found in GenBank. An app. 470 bp long region spanning from the end of exon 2 through the entire 2nd intron to the beginning of exon 3 was PCR-amplified and cloned. Clone sequences originating from a single specimen were firstly aligned by ClustalW in distinct matrices and every sequence type was retained in one representative/specimen. Retained sequences were then combined and aligned in a final matrix, on which phylogenetic tree reconstruction was based in Paup4.0b* using Maximum Parsimony criterion. Tree reconstruction worked with 561 characters of which 29 were parsimony-informative. Statistical support for branches was assessed via 1000 bootstrap pseudo-replication.

A heuristic search (TBR branch-swapping; AccTran optimization; gaps missing) found one

single most parsimonious tree at length 92 (CI = 0.9457, HI = 0.0543, RI = 0.9375), shown with bootstrap values in our figure. The two main subtrees has clearly shown similar topologies, the only difference between them was the lack of *A. albidus* sequences from one subtree.



The cloned LFY intron 2 sequences have provided a robust insight into the phylogeny of selected species of the *Astragalus* section Dissitiflori. It was evident that copies of the LFY gene belonged to two main gene-clusters representing either an ancient allopolyploidization or a duplication event. Notwithstanding, it has provided a clearer picture than nrITS, therefore, it can represent an alternative to nrITS in cases where the disadvantageous features of the latter challenge phylogenetic reconstruction.

This work was partially supported by a PhD scholarship (to L. Bartha) co-financed by the European Social Fund through the Sectoral Operational Program for Human Resources Development 2007-2013 (POSDRU/88/1.5/S/60185). The work of G. Sramkó was helped by the NKTH-OTKA-EU FP7 (Marie Curie action) co-funded 'MOBILITY' grant (nr: OTKA-MB08-A 80332).

Biomarkers of aging and aging-related pathologies

Moskalev AA[1], Batin MA[2]

[1]Institute of biology of Komi Science center of Ural division of RAS, [2]Moscow Institute of Physics and Technology;
mi20022@yandex.ru, amoskalev@list.ru

150,000 people die every day, two thirds of which die form age-related diseases. The main reasons of death are associated with cardiovascular diseases, diabetes, cancer, lung diseases and neurodegenerative pathologies. Any novel drug against age-related diseases has a huge market. However, there is no significant success in this area and the main reason for that is the lack of an exact system of effective biomarkers of aging and age-associated pathologies.

Any interventions (gene engineering, pharmacological, environmental) can't be used to the full extent in the absence of methods for evaluation of their effectiveness. Therefore, it is necessary to search for exact biomarkers of aging and methods to evaluate individual biological age, search for biomarkers to analyze epidemiological factors of age-related diseases.

Therefore, it is essential to systematization of potential biomarkers of aging, evaluation of their clinical effectiveness and development of mathematical tools for analyzing their impact.

The underlying processes of aging happen on the molecular (reactive oxygen species, protein cross-links, DNA strand breaks, epigenetic changes, endocrine shift), cellular (protein aggregation, mitochondrion dysfunction, lysosomal disfunction), tissue (cellular senescence, apoptosis, senescence-associated secretory phenotype) and systemic (immunosenescence, endocrinosenescence, chronic inflammation, regenerative capacity decreasing) levels. Changes associated with this phenomena (telomere truncation, advanced glycation end-products, amyloidosis, lipofuscin accumulation, NF-kB activity increasing, gene expression alteration) play a key role in development of age-related pathologies (cataract, senile systemic amyloidosis, Alzheimer's disease, cancer, brown atrophy of the heart, atherosclerosis, thrombosis, vascular dementia, heart failure, obesity, type 2 diabetes mellitus, arterial hypertension, chronic kidney failure, autoimmunity, sarcopenia, immunodeficiency) from one hand, and may serve as biomarkers of aging processes and effectiveness of geroprotective interventions on the other hand.

The report represents the chart of relationships between aging-associated phenomena, potential biomarkers of aging and age-related diseases.

Detecting past positive selection through ongoing negative selection

Georgii Bazykin[1], Alexey Kondrashov[2]

[1]*Institute for Information Transmission Problems of the Russian Academy of Sciences (Kharkevich Institute), Russian Federation, gbazykin@iitp.ru*
[2]*University of Michigan, United States, kondrash@umich.edu*

Detecting positive selection is a challenging task. We propose a method for detecting past positive selection through ongoing negative selection, based on comparison of the parameters of intraspecies polymorphism at functionally important and selectively neutral sites where a nucleotide substitution of the same kind occurred recently. Reduced occurrence of recently replaced ancestral alleles at functionally important sites indicates that negative selection currently acts against these alleles and, therefore, that their replacements were driven by positive selection. Application of this method to the Drosophila melanogaster lineage shows that the fraction of adaptive amino acid replacements remained ~0.5 for a long time. In the Homo sapiens lineage, however, this fraction drops from ~0.5 before the Ponginae-Homininae divergence to ~0 after it. The proposed method is based on essentially the same data as the McDonald-Kreitman test, but is free from some of its limitations, which may open new opportunities, especially when many genotypes within a species are known.

Analysis Of Genome-Wide Association Data In Chronic Sciatica Pain Cohort

Inna Belfer[1], Feng Dai[2]

[1]*Department of Anesthesiology, University of Pittsburgh, 3550 Terrace St, Pittsburgh, PA 15261, belferi@upmc.edu*
[2]*Yale Center for Analytical Sciences, 300 George Street, Suite 555, New Haven, CT 06511 feng.dai@yale.edu*

Aim of Investigation: Recent genome-wide association studies (GWAS) in humans have revealed novel genetic variants associated with disorders such as age-related macular degeneration, Crohn's disease, cancer, diabetes, and lipid disorders. However, over 440 recently published GWAS, only one study was related to human pain (acute post-surgical pain, [1]). To take advantage of recent advances in gene-mapping technology and hopefully fill the obvious

research-lag of pain genetics, we executed a genome-wide scan to identify genetic variants associated with chronic sciatica phenotype (summed pain score over one year after the surgery) in the sample of Maine patients [2].

Methods: First, genotypes determined with the Affymetrix SNP 500k chip were called using the Affy BRLMM algorithm. For data quality check, we removed individuals who have a genotyping rate of less than 95%, SNPs that have less than a 95% genotyping rate, SNPs that have a minor allele frequency (MAF) of less than 1%, or SNPs that fail the HW test with $p < 0.00001$. A total of 301286 SNPs were retained for final analysis of 177 patients with chronic sciatica pain.

The initial GWA data was analyzed for additive genetic model by fitting a regression model in which sex, age, SF36_GH, baseline worker's compensation status, and crossover to late surgery were included as covariates. Possible population stratification in patients was assessed by both genomic control and multidimensional scaling analysis using abundant genome-wide SNPs.

Results: Our initial analysis suggested that several SNPs were associated with chronic sciatic pain. Top five significant SNPs are rs4146308 (chr 15), rs48887298 (chr 15), rs10953178 (chr 7), rs71844628 (chr16) and rs11230889 (chr11) with a p-value at least less than 2.8 x 10-5. rs4146308 showed the strongest association with chronic pain ($p = 1.4 \times 10\text{-}6$). In Maine cohort, homozygotes for the rare "A" allele of this SNP (MAF = 0.18) had the higher pain score than homozygotes for the common "G" allele and heterozygotes "AG".

Perspective: More advanced analysis such as haplotype analysis, testing for gene x gene interaction, and copy-number variation analysis are being investigated by researchers; with replication analysis in other cohorts is currently being planned.

Acknowledgements: United States Cancer Pain Relief Committee. We also appreciate the great leadership from our late colleague Dr. Mitchell Max.

1. H. Kim, E. Ramsay, H. Lee, S. Wahl, R.A. Dionne (2009) Genome-wide association study of acute post-surgical pain in humans. *Pharmacogenomics*, **2:**171-179.
2 S.J. Atlas, R.A. Deyo, R.B. Keller, A.M. Chapin, D.L. Patrick, J.M. Long, D.E. Singer (1996) The Maine Lumbar Spine Study, Part II. 1-year outcomes of surgical and nonsurgical management of sciatica. *Spine*, **15**:1777-1786.

Systems Biology and Synthetic Bioengineering for Bioenergy Applications

Alexander S. Beliaev[1], Allan Konopka[1], Grigoriy Pinchuk[1], Mary Lipton[1], Thomas Squier[1], Aaron Wright[1], Thomas Metz[1], Jennifer Reed[2], Matthew Posewitz[3], and Donald Bryant[4]

[1]*Pacific Northwest National Laboratory*, Richland, WA, alex.beliaev@pnl.gov*
[2]*University of Wisconsin, Madison, WI*
[3]*Colorado School of Mines, Golden , CO*
[4]*The Pennsylvania State University, University Park, PA.*

The challenges to solar biofuel production are fundamental in nature and are associated with the inherent complexity of photosynthetic metabolism. To achieve the goal of economic, large-scale generation of biofuels, practical engineering and fundamental microbial physiology, biochemistry, and genomics must be reconciled to improve rates of photoautotrophic biomass productivity and the specific content of desired products. These challenges are fundamental and require profound knowledge of biological systems. This is where a viable solution will require a cross-cutting approach that not only leverages the availability of new genomic tools and engineering principles but also takes advantage of the systems biology suite of approaches to interrogate the genetic and functional diversity at an organism and community levels. It requires a shift of emphasis from "de novo" synthesis of complex functions to systems engineering, which will identify and manipulate the properties of a system as a whole and, which in a biological context, may greatly differ from the sum of the parts' properties. Our concept for developing viable technologies for bio-solar production of high-density fuels involves platforms beyond those used traditionally for metabolic engineering and synthetic biology (*i.e.*, *Escherichia coli*, *Saccharomyces cerevisiae*). The reason for this is relatively simple – engineering biological systems for environmental and industrial processes relevant to bioenergy missions will require biological systems with capabilities and adaptations that cannot be readily engineered into traditional platform organisms, *i.e.*, photosynthesis, extracellular electron transfer reactions, adaptation to extremes of temperature, salinity, pH, solvent concentrations, etc.

To that end, the U.S. Department of Energy Biofuels Scientific Focus Area (BSFA) at Pacific Northwest National Laboratory (PNNL) conducts fundamental research on cyanobacteria with specific emphasis on pathways of carbon, nitrogen, and redox metabolism that consume reductant and conserve energy produced by photosynthetic light reactions. The long-term objective of the PNNL BSFA is to develop a predictive understanding of metabolic subsystems

and regulatory networks involved in solar energy conversion to biofuel precursors or products. Toward this goal, we are conducting system-level analysis of modules involved in photosynthetic energy conservation and reductant generation; CO_2 accumulation, fixation, and reduction; biosynthesis of metabolic intermediates and monomers; and macromolecular synthesis. The gathered information is integrated into genome-scale metabolic reconstructions to understand metabolism qualitatively and quantitatively through a constraint-based flux analysis. When coupled with methods for the reconstruction of regulatory networks from transcriptomic data (and, as we look forward, proteomic data on post-translational modification of enzyme catalysts and transcription factors), this approach can provide experimentally testable predictions of: (*i*) fluxes to energy and reductant, (*ii*) reductant partitioning to carbon metabolism and other sinks, and (*iii*) analysis of anabolic and biosynthetic pathways that lead to macromolecular synthesis. The BSFA research embodies both scientific and technical tasks and builds on PNNL's expertise in systems biology, microbial physiology, genomics, proteomics, and metabolomics. This presentation discusses approaches for designing of novel pathways and cellular functions through genetic and metabolic engineering and summarizes the current progress in developing alternative engineering platforms for high-density fuels production.

*PNNL is operated for the DOE by Battelle Memorial Institute under Contract DE-AC06-76RLO 1830. Battelle Memorial Institute is ABO Corporate Member.

Evaluating Mixture Models for building RNA knowledge-based potentials

Adelene Y.L. Sim[1], Olivier Schwander[2], Michael Levitt[3], Julie Bernauer[4]

[1]*Department of Applied Physics, Stanford University, Stanford, CA, USA*
[2]*Combinatorial models team, Laboratoire d'Informatique (LIX), École Polytechnique, Palaiseau, France*
[3] *Department of Structural Biology, Stanford University, Stanford, CA, USA*
[4] *INRIA AMIB Bioinformatique, Laboratoire d'Informatique (LIX), École Polytechnique,Palaiseau, France*,
julie.bernauer@inria.fr

RNA molecules are involved in most biological processes and being able to predict their structure is key in the understanding of their function. Recent efforts in RNA structure prediction techniques have shown that parameterized energy functions and knowledge-based techniques largely improve the accuracy of structure prediction [1]. For protein folding, knowledge-based (KB) potentials have been developed [2, 3]. Most of these are based on Potential of Mean Force generated from distance distributions between atoms and can be used to score and refine protein structures. Due to the previous lack of available RNA 3D structures, KB potentials were just recently adapted to RNA molecules. We recently showed that they can be accurately used for screening both at atomic and coarse-grained representation levels [4].

To obtain a KB potential, a PMF is derived from distance measurements based on an energy function model first described by Samudrala and Moult [3]. The potential between atoms i and j is written as : $E = -kT\sum_{ij}\ln\left(P_{obs}(d_{ij})/P_{ref}(d_{ij})\right)$ where d_{ij} is the distance between i and j, T is the temperature (taken to be 300 K) and k the Boltzmann constant. $P_{obs}(d_{ij})$ and $P_{ref}(d_{ij})$ are the observed and reference probabilities respectively for the atoms i and j.

Obtaining, $P_{obs}(d_{ij})$ and $P_{ref}(d_{ij})$ is not an easy task. They are usually computed by binning distances. This binning calls for an interpolation function to evaluate the energy and be able to differentiate the energy function to perform refinement and molecular dynamics using KB potentials. The bin size has an effect on the prediction as described in [5].

In this study we show that the use of Gaussian Mixture Models (MM) allows an accurate description of $P_{obs}(d_{ij})$ and $P_{ref}(d_{ij})$. We can then derive efficient energy functions using various techniques: Kernel Density Estimation (Standard and Simplified), Expectation Maximization based and Dirichlet Process Mixture Models. Using an in-lab MM library [6], we compare the fitting to the data for these three techniques and show their performance for scoring RNA decoy

3D structures. We also show that using MM, it is possible to provide differentiable energy functions suitable for molecular dynamics experiments.

References

1. Das, R., J. Karanicolas, and D. Baker,. Nat Methods, 2010. **7**(4): p. 291-4.
2. Lu, H. and J. Skolnick, *A distance-dependent atomic knowledge-based potential for improved protein structure selection*. Proteins, 2001. **44**(3): p. 223-32.
3. Samudrala, R. and J. Moult, *An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction.* J Mol Biol, 1998. **275**(5): p. 895-916.
4. Bernauer, J., et al., *Fully differentiable coarse-grained and all-atom knowledge-based potentials for RNA structure evaluation.* RNA, in press.
5. Summa, C.M. and M. Levitt, *Near-native structure refinement using in vacuo energy minimization.* Proc Natl Acad Sci U S A, 2007. **104**(9): p. 3177-82.
6. Schwander, O. and F. Nielsen. *PyMEF - A framework for Exponential Families in Python*. in *IEEE Workshop on Statistical Signal Processing*. 2011.

Acknowledgments The authors thank the INRIA Équipe Associée program for financial support (GNAPI Associate Team). This work was supported by a National Institutes of Health award (GM041455) to M.L., a Human Frontiers in Science Program grant to M.L. A.Y.L.S. acknowledges support from the Agency for Science, Technology and Research (A*STAR), Singapore. The authors acknowledge support from NSF award CNS-0619926 for computer resources.

Predicting copy number alterations and structural variants using-paired end sequencing data

Valentina BOEVA, B. ZEITOUNI, K. BLEAKLEY, A. ZINOVYEV, J.-P. VERT, I. JANOUEIX-LEROSEY, O. DELATTRE

Institut Curie, 26 rue d'Ulm, Paris, F-75248 France; INSERM, U900, Paris, F-75248 France; Mines ParisTech, Fontainebleau, F-77300 France; INSERM, U830, Paris, F-75248 France; INRIA, Saclay, France; valentina.boeva@curie.fr

Emmanuel BARILLOT

emmanuel.barillot@curie.fr

The detection of structural variants (SVs) in the human genome plays an important role in the understanding of many genetic diseases, including cancer. In cancer, tumor suppressor genes can be deleted or mutated, whereas oncogenes can be amplified or mutated with a gain of function. Translocations can result in cancer-causing fusion proteins (BCR/ABL fusion in CML, BCL1/IGH in multiple myeloma, EWS/FLI1 in Ewing sarcoma, etc.)

With the arrival of new high-throughput sequencing technologies, our current power to detect SVs has significantly improved. Genomic breakpoints of large structural variants (i.e., translocations or large duplications and deletions) can be identified using two complementary approaches: calculation of copy number profiles (CNPs) and analysis of 'discordant' mate-paired/paired-ends mappings (PEMs).

Investigation of CNPs allows identification of genomic regions of gain and loss. There exist two frequent obstacles in the analysis of cancer genomes: absence of an appropriate control sample for normal tissue and possible polyploidy. We therefore developed a bioinformatics tool, called FREEC [1], able to automatically detect copy number alterations (CNAs) without use of a control dataset. FREEC normalizes copy number profiles using read mappability and GC-content and then applies a LASSO-based segmentation procedure to the normalized profiles to predict CNAs.

For PEM data, one can complement the information about CNAs (i.e., output of FREEC) with the predictions of structural variants (SVs) made by another tool that we have developed, SVDetect [2]. SVDetect finds clusters of 'discordant' PEMs and uses all the characteristics of reads inside the clusters (orientation, order and clone insert size) to identify SV type. SVDetect allows identification of a large spectrum of rearrangements including large insertions-deletions, duplications, inversions, insertions of genomic shards and balanced/unbalanced intra/inter-chromosomal translocations.

Here we present a package for automatic intersection of FREEC and SVDetect outputs that allows one to (1) refine coordinates of CNAs using PEM data and (2) improve confidence in calling true positive rearrangements (particularly, in ambiguous satellite/repetitive regions).

Both SVDetect and FREEC are compatible with the SAM alignment format and provide output files for graphical visualization of predicted genomic rearrangements.

Funding: The Ligue Nationale contre le Cancer (V.B., A.Z., E.B., I.J.-L. and O.D. are members of a labeled team).

1. V. Boeva et al. (2011) Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization, *Bioinformatics*, **27**(2):268-9.
2. B. Zeitouni et al. (2010) SVDetect - a bioinformatic tool to identify genomic structural variations from paired-end next-generation sequencing data, *Bioinformatics*, **26:** 1895-1896.

Phylogenomics and Robust Construction of Prokaryotic Evolutionary Trees

Katerina Korenblat[1], Zeev Volkovich[1], Alexander Bolshoy[2]

[1]*Software Engineering Department, ORT Braude Academic College, Karmiel, Israel,* katerina@braude.ac.il, vlvolkov@braude.ac.il
[2]*Department of Evolutionary and Environmental Biology, University of Haifa, Haifa 31905, Israel,* bolshoy@research.haifa.ac.il

Here we present a novel phylogenomic method of genome-tree construction on the basis of gene lengths of orthologous genes presented in completely sequenced genomes of prokaryotic organisms using Clusters of Orthologous Groups (**COG**s). Every single element of our input data is a median protein length related to a pair (**COG**, genome). In principle, the method is so fast that input data may consist of median protein lengths related to thousands of **COG**s and hundreds of genomes. Clustering is performed using an application of the information bottleneck method for unsupervised clustering.

Two main strategies in the field of a species tree phylogenomic reconstruction were developed to this end: the supertree and the supermatrix. One group of the supermatrix methods is associated with a Boolean matrix based on the presence and absence of gene families in genomes. Even though the method we present here is closely related to a group of methods based on the presence and absence of genes, it uses the information related to the lengths of genes, and this addition makes a significant difference.

In introducing a novel supermatrix phylogenomic method, we have had several primary goals.

- First, to propose a fast method that allows the use of whole proteome characters for reliable construction of genome trees. We show that the method is fast and reliable, indeed.
- Second, to show the robustness of the proposed algorithm. For robustness evaluation, we applied jackknife technique to input data. The aim of this approach is to show that tree structure based on different subsets of COGs is sufficiently stable. We have conducted extensive experiments to validate the performance of bootstrapping and jackknifing in order to estimate how robust the phylogenies produced by the proposed methodology are. These experiments show that randomization as part of the bootstrap procedure substantially decreases stability of the obtained trees and that jackknifing is very useful to determine the confidence level of a phylogeny.
- Thirdly, to reveal the phylogenetic nature of these trees on the basis of a few empirical case studies. We demonstrate that a selected small group of genomes is distributed reasonably along a produced phylogenetic tree. Although our comprehensive genome clustering is independent of phylogenies based on the level of homology of individual genes, it correlates well with the standard “tree of life" based on sequence similarity of 16s rRNA. This, together with successful jackknifing for the determination of confidence levels signifies that the method may be truly classified as phylogenomic. We have also examined several of the methodological issues involved in going from a large sequence database to a useable phylogeny. In particular, we integrate (semi) automated solutions to rogue taxon identifications and jackknifing measurements of tree stability in a single study to examine the phylogenetic signal contained in large sparse supermatrices.
- On the basis of a few empirical case studies, we intended to fix the parameters of the method. We considered three parameters to choose the most appropriate values of the parameters. (1) A bootstrapping parameter that designates a fraction of randomly resampling columns (COGs) of the input dataset. (2) A jackknifing parameter that designates a fraction of randomly deleted columns. (3) A preprocessing parameter (threshold) to consider only those columns of the supermatrix containing more elements than a certain threshold.

To summarize, we are confident in our proposal to construct prokaryotic phylogenetic trees using the fast and reliable method described in this manuscript with parameter values equal to 15% of the maximal COG size for the preprocessing parameter and equal to 80% for the jackknifing parameter.

Average-case analysis methods dedicated to the study of Biological Networks

Jeremie Bourdon

LINA CNRS UMR 6241, University of Nantes, France, Jeremie.Bourdon@univ-nantes.fr

Despite recent improvements in molecular techniques, biological knowledge remains incomplete.

Any theorizing about living systems is therefore necessarily based on the use of heterogeneous and partial information. Much current research has focussed successfully on the qualitative behaviors of macromolecular networks. Nonetheless, it is not able to take available quantitative information such as time-series protein concentration variations into account.

The present work proposes a probabilistic modeling framework that integrates both kinds of information. Average case analysis methods are used in association to Markov chains in order to link qualitative information about transcriptional regulations to quantitative information about protein concentrations. The approach is illustrated by modeling the carbon starvation response in Escherichia coli.

Its use accurately predicts the quantitative time-series evolution of several protein concentrations by using only discrete gene interaction knowledge and very few quantitative observation on a single protein concentration. From that, the modeling technique also derives a ranking of interactions with respect to their importance during the considered experiment. Such a classification is confirmed by literature knowledge. Therefore, as a main novelty, our method permits (i) to integrate few quantitative information into an existing qualitative discrete model and derive new quantitative predictions, (ii) to precisely quantify the robustness and relevance of interactions with respect to phenotypic criteria, and (iii) to extract the key features of the model and design some new experiments.

Physicochemical and structural properties determining HIV-1 coreceptor usage

Kasia Bozek[1], Thomas Lengauer[1], Francisco Domingues[2]

[1]*Max Planck Institute for Computer Sciences, Germany, bozek@mpi-inf.mpg.de*
[2]*EURAC Institute of Genetic Medicine, Bozen, Italy*

The entry of the human immunodeficiency virus (HIV) into human cells is a multi step process involving binding to one of the cell-surface coreceptors CCR5 or CXCR4 (1). The binding site of the coreceptor is partially located in the third variable region (V3) of gp120 viral protein (2). Whether a virus can bind to CCR5 only (R5 virus), to CCR5 and CXCR4 alternately (dual virus) or to CXCR4 only (X4 virus) is determined predominantly by the sequence and structure of this region. The phenotype related to the virus coreceptor usage is termed viral tropism. While in the early, asymptomatic stages of infection mainly R5 viruses are observed, progression to AIDS is often correlated with the emergence of X4 viruses (3). The relationship of HIV tropism with disease progression and the recent development of CCR5-blocking drugs underscore the importance of monitoring virus coreceptor usage. As an alternative to costly phenotypic assays, computational methods aim at predicting virus tropism based on the V3 loop sequence and on its structure (4-7). The major drawback of the binary sequence representation is the limited insights it offers into the physicochemical properties of amino acids and their spatial arrangement in the binding site that determine coreceptor binding.

Here we present a structural descriptor of the V3 loop encoding the physicochemical properties of the loop together with their locations on the protein structure. We map 54 amino acid indices representing the physicochemical properties of amino acids onto the V3 loop structure and use machine learning methods to extract the features which are the most informative for coreceptor usage. The extracted set of features represents a small fraction of the initial feature set and models based on this set attain higher prediction accuracy with decreased computational load.

Our descriptor used as input to the support vector machine (SVM) predicting tropism shows a statistically significant improvement over the binary representation of the V3 sequence. At the specificity of 11/25 rule a sensitivity of 69% was achieved, comparing favorably with the 62% sensitivity of sequence-based prediction. In addition to the data inferred from lab-cloned viruses (clonal data) we assessed the predictive power of our method on the clinically derived 'bulk' sequence data of patient samples and obtained a statistically significant 3% improvement

over the sequence representation evaluated using receiver operating characteristic (ROC) curve. We also demonstrated the capacity of our method to predict the outcome of therapies based on coreceptor blocker Maraviroc (8) by applying it to 53 samples of patients undergoing Maraviroc therapy.

Our structural descriptor affords direct interpretation of the features of the V3 loop relevant for viral tropism by indicating specific physicochemical properties of amino acids in distinct parts of the loop being predictive of coreceptor usage. The analysis of features important for the classification pointed to two loop regions playing determining role in the coreceptor usage. The regions are located on the opposite strands of the loop stem and show predominantly structure, hydrophobicity and charge-related properties. In the bound conformation of the loop these regions are situated in close proximity forming a potentially determinant site for the coreceptor binding.

The proposed method offers therefore higher performance over sequence-based method with a comparable efficiency and a direct interpretation of structural and physicochemical determinants of tropism.

1. Chan, D. C., and P. S. Kim. (1998) HIV entry and its inhibition. Cell 93:681-4
2. Fouchier, R. A., M. Groenink, N. A. Kootstra, M. Tersmette, H. G. Huisman, F. Miedema, and H. Schuitemaker. (1992) Phenotype-associated sequence variation in the third variable domain of the human immunodeficiency virus type 1 gp120 molecule. J Virol 66:3183-7.
3. Berger, E. A., P. M. Murphy, and J. M. Farber. (1999) Chemokine receptors as HIV-1 coreceptors: roles in viral entry, tropism, and disease. Annu Rev Immunol 17:657-700.
4. Jensen, M. A., F. S. Li, A. B. van 't Wout, D. C. Nickle, D. Shriner, H. X. He, S. McLaughlin, R. Shankarappa, J. B. Margolick, and J. I. Mullins. (2003) Improved coreceptor usage prediction and genotypic monitoring of R5-to-X4 transition by motif analysis of human immunodeficiency virus type 1 env V3 loop sequences. J Virol 77:13376-88
5. Sing, T., A. J. Low, N. Beerenwinkel, O. Sander, P. K. Cheung, F. S. Domingues, J. Buch, M. Daumer, R. Kaiser, T. Lengauer, and P. R. Harrigan. (2007) Predicting HIV coreceptor usage on the basis of genetic and clinical covariates. Antivir Ther 12:1097-106.
6. Sander, O., T. Sing, I. Sommer, A. J. Low, P. K. Cheung, P. R. Harrigan, T. Lengauer, and F. S. Domingues. (2007) Structural descriptors of gp120 V3 loop for the prediction of HIV-1 coreceptor usage. PLoS Comput Biol 3:e58.
7. Dybowski, J. N., D. Heider, and D. Hoffmann. (2010) Prediction of co-receptor usage of HIV-1 from genotype. PLoS Comput Biol 6:e1000743.
8. Dorr, P., M. Westby, S. Dobbs, P. Griffin, B. Irvine, M. Macartney, J. Mori, G. Rickett, C. Smith-Burchnell, C. Napier, R. Webster, D. Armour, D. Price, B. Stammen, A. Wood, and M. Perros. (2005) Maraviroc (UK-427,857), a potent, orally bioavailable, and selective small-molecule inhibitor of chemokine receptor CCR5 with broad-spectrum anti-human immunodeficiency virus type 1 activity. Antimicrob Agents Chemother 49:4721-32.

Inference of ancestral states for character evolution: the case of uncertain data at terminal nodes

Nadezda Bykova, Andrey Mironov

Dept. of Bioengineering and Bioinformatics, M.V. Lomonosov Moscow State University, Vorobievy Gory 1-73, Moscow 119991, nadelle4@mail.ru

The problem of reconstructing ancestral states given a phylogeny and data from extant species arises in many areas of bioinformatics. Commonly used technique is the Markovian probabilistic model, where the rates of transitions between states are defined relatively to evolution rates. Maximizing the probability to observe given data one can find optimal rate parameters, and then reconstruct ancestral states by their contribution in overall probability. Recent additions to this method were developed account for uncertainties in the initial phylogeny [1]. Here we modify the basic Markovian model to apply in a situation when one has not exact data, but predictions about states at terminal nodes.

We aim to reconstruct the ancestral states given the tree phylogeny, prediction scores at terminal nodes and prediction score distribution for different states. This probability distribution could be collected from a large unbiased sample of predictions. To reconstruct the optimal set of ancestral states we use the Viterbi algorithm, modified for a tree:

$$\begin{cases} V_Y^i = \max_j(P^{ij}(t_M) * V_M^j) * \max_k(P^{ik}(t_L) * V_L^k), \ when\ \{M,L\} = child(Y) \\ V_Y^i = \rho^i(x_Y), \text{when } \emptyset = \text{child(Y)} \end{cases},$$

where ρ^i(x) is the score probability distribution for state *i*; $P^{ij}(t)$ is the probability of transition from state *i* to *j*; t_L is the distance from node Y to the ancestral node; *child(Y)* is the set of descendants of node Y;

As it the linear Viterbi algorithm, after selecting the most probable state at the root, the states of descendant nodes can be reconstructed at the reverse run as argmax.

To reconstruct the state probabilities at each node we use the forward-backward algorithm, modified for a tree:

$$\begin{cases} B_Y^i = \sum_j \left(P^{ij}(t_M) * B_M^j\right) * \sum_k \left(P^{ik}(t_L) * B_L^k\right), when\ \{M,L\} = child(Y) \\ B_Y^i = \rho^i(x_Y), when\ \emptyset = child(Y) \end{cases}$$

$$\begin{cases} F_Y^i = \sum_j (F_Z^j * P^{ji}(t_Y) * \sum_k (P^{jk}(t_W) * B_W^k)), when\ W = neighb(Y) \\ F_Y^i = \omega(i), when\ \emptyset = neighb(Y) \end{cases}$$

where $\omega(i)$ is the prior probability of states at the root; *neighb(Y)* is the neighbor node to node Y.

The final state probability at a node can be found after the forward and backward runs as:

$$P_Y^i = \frac{F_Y^i * B_Y^i}{\sum_i (B_{root}^i * \omega(i))}$$

In fact, similar algorithms were proposed for defined discrete states at terminal nodes: the Viterbi algorithm was described as "weighted parsimony", the backward algorithm was described as the "purning" algorithm by Felsenstein [2]. The forward recursive formulas for a tree were not published, although a hint was given in [1]. With the addition of emission probabilities at terminal nodes, the reconstruction of ancestral states becomes closer to the HMM (Hidden Markov Model) reconstitution. Hence we believe that using the HMM language could be helpful in further development of this model (for example, using more complex HMM structures to account for parameters changing between nodes).

It should be also noted, that when we reconstruct the state probabilities at terminal nodes, we obtain the posterior probabilities of states in the extant species (prior probabilities could be calculated from the prediction scores and score distributions). So the method could be considered as a novel comparative genomic technique to correct predictions, where the correspondence of sequence similarity to feature similarity is defined by the optimized transition rates for different states. The species hierarchy is also taken into account, rather than the distances only.

The method is illustrated with an example of predicted N-terminal signal peptides in bacterial proteins from ortholog clusters.

Implementing the opportunity to reconstruct the evolutionary history from predicted data would expand applicability of the Markovian probabalistic method to new areas of bioinformatics, such as predictions of molecular traits.

This study was partially supported by RFBR grant 09-04-92742.

1. M.Pagel et al. (2004) Bayesian estimation of ancestral character states on phylogenies, *Syst Biol*, **53(5)**:673-84.
2. J. Felsenstein (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach, *J Mol Evol*, **17(6)**:368-76.

Protein-protein interfaces – structural features, and changes brought about by complex formation

Pinak Chakrabarti

Department of Biochemistry and Bioinformatics Centre, Bose Institute, Kolkata 700054, India India,
pinak@boseinst.ernet.in

To understand the physicochemical features underlying molecular recognition structural databases have been generated representing different types of protein-protein interfactions, such as 'strong' homodimers (which are permanent in nature), 'weak' dimers (that can exist in equilibrium between monomer and dimer), transient protein-protein heterocomplexes, and the non-physiological interfaces that are observed in the crystal lattice of protein crystals. These categories do show some differences in the atomic packing [1]. The average interface area (the accessible surface area, ASA, on the two components that gets buried on heterocomplex formation) is ~1600 Å2 and the interface can be dissected into core and rim regions with the former being composed of residues that are more conserved than those in the latter [2].

To gain an insight into the process of molecular recognition and protein-protein interaction one needs to understand not only the static interface formed between the two molecules, but also the changes that result in the structure as the two components associate. To analyze the physicochemical changes that occur in an isolated protein structure as it forms a complex, we have used the Protein-Protein Docking Benchmark (version 3.0) [3] and mapped the interface residues and atoms as seen in the complex to those in the isolated state. On going from the isolated (I) to the complex (C) state there is a change of about 6% in the accessible surface area (ASA) of the interface atoms (considering the absolute value of the difference and relative to the total value in the complex state). Other changes, especially those occurring in the secondary structural content would be discussed.

[1] Dey S, Pal A, Chakrabarti P and Janin, J. The subunit interfaces of weakly associated homodimeric proteins. J. Mol. Biol. 398, 146-160 (2010)
[2] Guharoy, M. and Chakrabarti, P. Conservation and relative importance of residues across protein-protein interfaces. Proc Natl Aca. Sci USA 102, 15447-15452 (2005).
[3] Hwang H., et al . (2008) Protein-protein docking benchmark version 3.0, Proteins; 73:705-709.

Profile periodicity of DNA coding regions

Maria Chaley[1], Vladimir Kutyrkin[2]

[1] *Institute of Mathematical Problems of Biology, Russian Academy of Sciences, Institutskaya st., 4, 142290, Pushchino, Russia, maramaria@yandex.ru*
[2] *Department of Computational Mathematics and Mathematical Physics, Moscow State Technical University n.a. N.E. Bauman, the 2nd Baumanskaya st., 5, 105005, Moscow, Russia*

Research on both short- and long-range nucleotide correlations is of great importance for understanding known structural particularities in DNA sequences and for revealing new ones. Images of various functions demonstrating nucleotide correlations in DNA coding regions show regular peaks with the steps of three bases corresponding to the triplet nature of the genetic code. This has led to the notion of triplet periodicity in the coding regions.

In light of current understanding of latent periodicity as approximate tandem repeats [1], the occurrence of periodicity has been corroborated by textual "consensus-pattern", which is an estimate of a pattern in the original repeat. If alterations in the copies of the pattern account for more than 30% of the pattern, the validity of the revealed consensus-pattern is in doubt. Although tandem repeats of three- and hexa-nucleotides occur in coding regions, as a rule it is impossible to deduce a reliable consensus-pattern of approximate tandem repeat over the whole length of a coding region.

Use of Fourier methods for revealing imperfect or latent periodicity has become common. Other statistical methods have also arisen for determining latent periodicity in nucleotide sequences. These methods are based upon measuring heterogeneity in nucleotide distributions over period positions. In practice, in the absence of weak periodicity in a sequence that is not an approximate tandem repeat, a high index of heterogeneity and a Fourier spectrum with a dominant peak may be observed. Nevertheless, it is incorrect in this case to use the term "latent periodicity", until the discovery of a pattern indicating some new type of periodicity.

A spectral-statistical approach [3] to identifying a new type of latent periodicity, called profile periodicity or profility, is presented. The notion of latent profility in DNA sequences [2] expands on the idea of approximate tandem repeat [1] in which textual string (DNA sequence) is presented as a chain of eroded copies (with approximately 80% identity) of the textual pattern. Latent profile periodicity occurs in DNA regions where nucleotide correlations can be described by hypothesizing on the generation of successive uni-length DNA fragments according to a fixed probability distribution of nucleotides appearance at each fragment position. A pattern of latent profility can be described with the aid of a finite random string consisting of independent random

characters with corresponding probability distribution for the textual characters from the DNA alphabet. The usual methods [1] applied for identification of approximate tandem repeats therefore cannot be used to reveal new types of latent periodicity.

Via a procedure that identifies patterns of latent profile periodicity in DNA [3], it is possible to reveal the two levels of organization in encoded genetic information: regular heterogeneity of nucleotide distribution over positions in the codons, and latent profility. It is shown that Fourier analysis does not enable the second level (latent profility) to be distinguished in the genetic encoding. The latent profility revealed in the DNA coding regions can be translated into features of protein structure. It is possible to determine signs of local features in the structure of genes and corresponding proteins through latent profile periodicity (latent profility). The direct detection of such features is challenging because the goal of the search is not a priori known.

1. G. Benson (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucl. Acids Res.*, **27:** 573-580.
2. 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, *Math. Biosci.*, **211:** 186-204.
3. M.Chaley, V. Kutyrkin (2010) Structure of proteins and latent periodicity in their genes. *Moscow Univ. Biol. Sci. Bull.*, **65:** 133–135.

Significance of Clusterin Expression in Patients with Hepatocellular Carcinoma Undergoing Hepatic Resection

Gar-Yang Chau

Taipei Veterans General Hospital, Taiwan, gychau@vghtpe.gov.tw

Background: Surgical resection offers hepatocellular carcinoma (HCC) patients a chance for a cure, but carries a high tumor recurrence rate. Clusterin is a highly conserved glycoprotein, enhancing cell aggregation in vitro. Clusterin upregulated in cancers of the breast, ovary, colon, prostate, kidney and HCC. Overexpression of clusterin has been correlated with tumor metastasis. This study evaluates the significance of serum clusterin levels and protein expression in resected tissue specimen in HCC patients.

Methods: The sera and HCC and nontumor tissues of 140 HCC patients undergoing hepatic resection were prospectively collected. Serum clusterin levels were determined by enzyme-linked immunosorbent assay. The clusterin protein expression in resected specimen were examined by immunohistochemistry. Median followup time was 39.1 months.

Results: Mean serum clusterin levels were 13.0 ± 5.9 ng/mL (range, 1.0~36.7 ng/mL). Patients with high serum clusterin level (> 14.5 ng/mL) had significantly lower frequency of family history of HCC, poorer liver function, and higher frequency of tumor being multiple, presence of vascular invasion than those with low clusterin level (≤ 14.5 ng/mL). For patients with tumor size > 5 cm, patient with a high serum clusterin level had significantly worse postoperative survival when compared to patients with low serum clusterin level (p = 0.018). Clusterin protein was overexpressed in HCC tissues in 76 patients (54.3%) and in nontumor liver tissues in 53 patients (37.9%). Patients with overexpression of clusterin in nontumor tissue have worse postoperative survival rates (p=0.003).

Conclusions: In HCC patients, high serum clusterin levels and overexpression of clusterin in nontumor tissue related with worse prognosis after hepatic resection. Clusterin can be a prognostic indicator for HCC patients after resection.

3-D complexes of VirE2 protein originated from Agrobacterium tumefaciens and evaluation of his pore-forming ability

Mikhail Chumakov[1,2], Yurii Gusev[1,2], Svyatoslav Mazilov[2]

[1]*Institute of Biochemistry and Physiology of Plants and Microorganisms, Russian Academy of Sciences, 13 Prospekt Entuziastov, Saratov 410049, Russia*
[2]*Bioph.Dept. Non-linear Processes Faculty, Saratov State University, Saratov 410025, Russia*
chumakov@ibppm.sgu.ru

Bacteria of the genus *Agrobacterium* are a natural vector for the transfer of genetic information (T-DNA) into the eukaryotic cell. The T-DNA nonspecifically integrates into the host's chromosome and is inherited at subsequent cell divisions. The T-DNA–VirD2 complex and VirE2 proteins are transferred to the host cell independently. It is belived that VirE2 proteins form a membrane-spanning pore or a channel for promotion of short nucleotides translocation across the membrane [1]. How VirE2 participates in the T-DNA transfer across artificial and natural membranes is not known.

For prediction of secondary structures for VirE2, the PROFsec program, located at http://www.embl-heidelberg.de, was used. For formation of a complex consisting of VirE2 protein subunits, the GRAMM-X program, located at (http://vakser. bioinformatics.ku.edu), was used. For localization of VirE2-dependent complexes in lipid membranes, the CHARMM GUI-Membrane Builder program, located at (http://www.charmm-gui.org), was used.

Using the PROFsec program, we predicted 15 β-sheets [2], and 22 α-helices for the VirE2 fragment (amino acid residues 112 through 517). X-ray diffraction analysis showed that the same fragment of VirE2 as a complex with VirE1 (PDB ID:3BTP) contains 21 β-sheets and 15 α-helices [3]. Only in 5% of sequences did a contradiction arise in our prediction of helix–sheet structures: the presumed α-helix (amino acid residues 243 through 253) was predicted to be in the place of β-sheets (amino acid residues 242 through 246 and 249 through 251), shown by X-ray diffraction analysis. It follows that the results obtained with the PROFsec program may give grounds to make suggestions as to the secondary structure of proteins that lack like VirE2 homology to known proteins. Duckely and Hohn [1] hypothesized that the β-barrel fold is a structural solution for the delicate transition from soluble to membrane-bound state and possibly for the passage of ssDNA through membranes. Subsequently, it was found that VirE2 forms a TIM-like barrel composed of α-helices and β-strands, resulting in a donut shape with interior β-strands and exterior α-helices [1]. We found that the TIM-like barrel located in the C-domain of VirE2 had a channel. We measured the inner diameter of the TIM-like barrel located in the C-domain of VirE2 and evaluated it to be 1.3 nm. Possibly, it is sufficient for the passage of short oligonucleotides through the pores formed by two VirE2 proteins in the bilayer membrane. However, the TIM-like barrel complex is not sufficient for the translocation of the ssT-DNA–VirD2 complex across the TIM-like barrel, since the VirD2 protein has a size of 2×4 nm, according to our evaluation of a 3D model for VirD2 by using the Swiss-PdbViewer program. In a probabilistic model by using the Membrane Builder program for a structure composed of two VirE2 proteins integrated into a lipid bilayer, a channel with a diameter of 2 nm may form between Ile330 VirE2 protein (number 1) and Ile330 VirE2 protein (number 2). One model structure built from four of VirE2 proteins by using the CHARMM GUI program had a channel (3.8 nm in size) between Lys488 VirE2 protein (number 1) and Lys488 VirE2 protein (number 3) in the membrane portion of the complex.

1. M. Duckely, B. Hohn (2003) The VirE2 protein of *Agrobacterium tumefaciens*: the Yin and Yang of T-DNA transfer. *FEMS Microbiol. Lett.* **223**:1–6.
2. M.I.Chumakov, A.V.Burmatov, V.I. Bogdanov, I.V. Volokhina (2004) Experimental and computer evaluation of the ability ssT-DNA-binding VirE2 protein to interact with lipid membranes. In.: Proc. The Fourth International Conference on Bioinformatics of Genome Regulation and Structure (BGRS'2004), Novosibirsk, Russia, July 25-30, V.1. P. 252-254.
3. O. Dym, S. Albeck, T. Unger, J. Jacobovitch, A. Branzburg et al. (2008) Crystal structure of the *Agrobacterium* virulence complex VirE1-VirE2 reveals a flexible protein that can accommodate different partners. *Proc. Natl. Acad. Sci. USA.* **105**:11170–11175.

Homo sapiens L. – a Species is in a State of Evolutionary Saltation

Vladimir Chupov, Eduard Machs

Komarov Botanical Institute of the Russian Academy of Sciences, Russian Federation, nika-egida@mail.ru

Motivation and Aim. We studied the process of macro-evolution in Angiosperms. It was revealed the saltational nature of this process and positive correlation between the evolutionary events, the level of evolutionary advance and the nucleotide composition, dinucleotide and CpG content in ITS1-5.8S-ITS2 rDNA [1-3].

The revealed correlation allows determine the evolutionary status of taxa according to the analysis of nucleotide composition of rDNA. On the assumption of this idea a problem of evolutionary status of the species Homo sapiens L. brought up.

Methods and Algorithms. We used the NCBI GeneBank data on ITS1, ITS2 and 5.8S rRNA sequences and MEGA-4 and DAMBE software.

Results. We consider our conclusions as preliminary because the rDNA of animals is less studied as compared with rDNA of plants. We analyzed various regions of rDNA of Homo sapiens, Primates (11 specirs), Rodentia (7 sp.), Carnivora (8 sp.), Reptilia (4 sp.) and Amphibia (5 sp). It was shown that both C+G and CpG content in rDNA of Homo sapiens, on a level with Pongo pigmaeus, is highest among the primates and even most saturated by guanine and cytosine among animals in general. The dinucleotide spectra of rDNA of man and apes relate to the most specialized groups containing maximum GG, CC, CpG and CpC and minimum quantity TpG dinucleotide.

Conclusion. These data being compared with the nucleotide spectra of plant rDNA in evolutionary advanced and static taxa bring to the conclusion that the evolutionary state of the species Homo sapiens correspond to the rarely observed stage of evolutionary saltation[5].

Гранты: "Динамика генофондов" и РФФИ 09-04-014-69, 09-04-01469-а.

1. Chupov, V. S. (2002) Shape of lateral phylogenetic branch in plants by data of neontological-taxonomic analysis of evolution. Uspechy sovremennoi biologii 122, 227-238. (In Russ., Enl. Summary)
2. Chupov, V. S., Punina, E. O., Machs, E. M., Rodionov, A. V. (2007) Nucleotide composition and contents of CpG in rRNA of the representatives of Melanthiales-Liliales and Melanthiales-Asparagales reflects a characteristic feature of evolution. Mol. Biol. (Moscow), 41: 808-829.
3. Chupov, V. S., Machs, E. M., Rodionov, A. V. (2008) The dinucleotide composition of rhibosomal spacer region ITSI-5.8S rDNA-ITS2 as an indicator of evolutionary development and a phylogenetic marker of monocotyledon plants Melanthiales - Liliales and Melanthiales - Asparagales (Angiospermae, Monocotyledones). General changes in the dinucleotide composition. Uspechy sovremennoi biologii 128, 481-496. (In Russ, Engl. Summary)
4. Chupov, V. S., Machs, E. M., Rodionov, A. V. (2008)The dinucleotide composition of the rDNA region (ITS1-5.8S-ITS2) as an indicator of evolutionary levels of development and phylogenetic marker of monocotyledon macrotaxa. dinucleotide spectrum of cryptaffine taxa
Uspechy sovremennoi biologii 128, 542–552. (In Russ, Engl. Summary)
5. Chupov, V. S. (2010) Специфические эволюционные особенности биологического вида Homo sapiens L. Футурологический конгресс: будущее России и мира.Материалы всероссийской научной конференции, 4 июня. 2010 г., Москва. С. 247-257.

Ribosomal multicopy protein study

Iakov Davydov[1], Irena Artamonova[2], Alex Tonevitsky[3]

[1]*Russian Research institute of physical education and sport, Russian Federation, davydov@bioinf.ru*
[2]*Vavilov Institute of General Genetics RAS, Russian Federation, irenart@vigg.ru*
[3]*Moscow State University, Russian Federation, tonevitsky@mail.ru*

L12 is the only protein present in prokaryotic ribosome in multiple copies. It is known that L12 interacts with translation factors such as EF-G, EF-Tu and RF3. While interacting with these factors L12 seems to increase their GTPase activity. Probably L12 also has other functions, e.g. it was recently shown that mitochondrial L12 controls the mitochondrial genes transcription rate [1].

L12 stalk comprised of several L12 copies is one of the least studied ribosome functional centers. Partially that is due to its high flexibility which complicates X-ray study [2].

In the end of 1970's it was shown that *Escherichia coli* ribosome contains four L12 molecules [3]. However stoichiometry 4:1 is not the only one possible option for the bacteria. *T. maritima* ribosome comprises six L12 molecules [2] as well as *Thermus themophilus* ribosome [4].

Our goal was to determine L12 stoichiometry for different species using genome sequence and to study the changes of L12 copy number between sibling species.

We developed an algorithm for L12 copy-number prediction. Our approach was based on the prediction of L10 secondary structure in order to locate L12 binding region. We analyzed several hundreds of bacteria and demonstrated that both variants (4:1 and 6:1) were prevalent among different phyla. It should be noted that mitochondria ribosome has 6:1 stoichiometry.

Analyzing the examples of different L12 copy numbers between closely related species we concluded that the reduction of L12 copy number was usually caused by the loss of a short gene fragment inside the eighth alpha helix of L10. The insertion of such a fragment could be resulted in the possibility to bind an additional L12 dimer, but it was observed much rarer.

It was also found that some cyanobacteria (such as Arthrospira and Gloeobacter) possibly had stoichiometry 8:1.This possibility was not noted previously.

The authors are supported by the Russian Ministry of Science grant 16.740.11.0449.Some of the results were obtained using the supercomputer SKIF MSU "Chebyshev" of the Moscow State University Supercomputer Center.

1. Z.Wang et al. (2007) Human mitochondrial ribosomal protein MRPL12 interacts directly with mitochondrial RNA polymerase to modulate mitochondrial gene expression, *J. Biol. Chem*, **282:**12610–12618.
2. M.Diaconu et al. (2005) Structural basis for the function of the ribosomal L7/12 stalk in factor binding and GTPase activation, *Cell*, **121:**991–1004
3. A.T.Gudkov et al. (1978) Stoichiometry and properties of the complex between ribosomal proteins L7 and L10 in solution, *FEBS Letters*, **93:**215–218.
4. L.Ilag et al. (2005) Heptameric (L12)6/L10 rather than canonical pentameric complexes are found by tandem MS of intact ribosomes from thermophilic bacteria, *Proc. Natl. Acad. Sci. U.S.A.*, **102:**8192–7.

An amino acid polymorphism centric view of classical HLA associations in complex traits

Paul de Bakker

Harvard Medical School, United States,
pdebakker@rics.bwh.harvard.edu

The major histocompatibility complex (MHC) region on chromosome 6 harbors the largest number of validated genetic associations to human diseases. The extreme genetic diversity of this locus and the broad linkage disequilibrium among variants has made it challenging to localize association signals and to pinpoint causal variants. Using specific examples of HIV control and autoimmunity, I will illustrate how association signals of common SNPs can be interpreted in terms of amino acid polymorphisms within classical HLA proteins.

Characterising Selection in human Conserved Non-coding Elements (CNEs) from the HapMap and 1000 Genomes Projects

Dilrini De Silva[1], Richard Nichols[1], Greg Elgar[2]

[1]*Queen Mary University of London, United Kingdom, d.desilva@qmul.ac.uk*
[2]*National Institute for Medical Research, United Kingdom, gelgar@nimr.mrc.ac.uk*

Conserved Noncoding Elements (CNEs) are a set of approximately 7000 conserved non-coding regions of the genome identified by whole genome comparison between human and *Takifugu rubripes* (Japanese Pufferfish), a vertebrate species evolutionarily distant from humans [5]. Arranged in clusters around genes regulating transcription and development (trans-dev genes) in vertebrates, some CNEs act as tissue-specific enhancers during embryonic development and are candidates for a network of *cis*-regulatory modules [5]. It has been suggested that a deleterious effect can be observed when most *cis*-regulatory elements which have one or more functions in the genome are lost [4]. We also know of cases where mutations in CNEs (not the gene itself) can lead to congenital abnormalities in human beings, for example, certain kinds of polydactyly [3] and the Pierre Robin syndrome [2] are caused by mutations in CNEs. Their absence in invertebrates suggests they may have played an important role in vertebrate development and evolution.

Efforts to catalogue human genetic variation by the HapMap and 1000 Genomes Projects have allowed us to analyse single nucleotide polymorphisms across worldwide populations in order to detect signals of selection in CNEs. Preliminary analysis of the allele frequency spectrum in CNEs shows a striking similarity to the skewed distribution observed at nonsynonymous sites indicating a signal of selection. Further analysis of genotype data from HapMap and 1000 Genomes Projects on the LOSITAN [1] workbench is expected to identify candidate loci under selection, which can then be used to study expression patterns in zebrafish.

1. Antao T. et al. (2008) LOSITAN: A workbench to detect molecular adaptation based on a Fst-outlier method, *BMC Bioinformatics*, **9**:323.
2. Benko S. et al. (2009) Highly conserved non-coding elements on either side of SOX9 associated with Pierre Robin sequence, *Nat Genet*. **41**: 359-364.
3. Lettice L. et al. (2003) A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum Mol Genet*. **12**: 1725-1735.
4. McLean C. and Bejerano G. (2008) Dispensability of mammalian DNA, *Genome Res*., **18**: 1743-1751.
5. Woolfe A. et al. (2004) Highly Conserved Non-coding Sequences Are Associated with Vertebrate Development. *PLoS Biol*. **3** (1): e7.

New benzimidazole derivatives as possible antibacterial drugs

Oleg Demchuk, Dmitro Lytvyn, Alla Yemets, Pavel Karpov, Yaroslav Blume
Institute of Food Biotechnology and Genomics, Natl. Acad. Sci. of Ukraine, Ukraine, demom79@gmail.com

The bacterial FtsZ proteins represent potential cytoskeletal target for antimicrobial compounds as essential component of bacterial cell division machinery. They are conserved in most important bacterial pathogens and responsible to target-based drug discovery toolbox with features of an ideal antimicrobial target [1, 2]. Due to homology with eukaryotic tubulins the benzimidazoles and their derivatives are considered as potential inhibitors of bacterial FtsZ proteins [3, 4]. Recently the libraries of novel tri-substituted benzimidazoles were created through rational drug design. The transmission electron microscopy and scanning electron microscopy analyses of *Micobacterium tuberculosis* FtsZ and bacterial cells, respectively, treated with a lead compound strongly suggest that benzimidazoles have a novel mechanism of action on the inhibition of mycobacterial FtsZ assembly and Z-ring formation [5].

According to data mentioned above, we investigated inhibition properties of eight new phenyl-thiazine derivatives of benzimidazole on functionality of bacterial FtsZ. Potential binding sites for them were identified on molecular surface of FtsZ by HEX 6.1 [6] with CUDA-graphics. For this study, we applied rigid docking of these compounds and reconstructed model of bacterial FtsZ as a target. The seven tested compounds had high scored binding site identical for 3-methoxybenzamide derivatives in region of cleft between helix seven (H7) and the C-terminal domain [7, 8]. Five of the tested compounds had lower score to nucleotide binding site. The predicted binding sites for the last benzimidazole derivative were localized in GTP binding site and along the surface of FtsZ between this site and structural T-7 loop.

We also estimated the FtsZ polymerization efficiency of these compounds *in vitro* by spectrophotometery. The solution density strongly correlated with light scattering of polymerized protein. Our study demonstrates the influence of all tested compounds in the first few minutes of reaction on the polymerization process. It was identified that two benzimidazole derivatives disturbed polymerization very sufficientle. Interestingly, the way action of these two compounds was opposite: the first inhibited polymerization and the second inhibited filament depolymerization. Since normal functioning of FtsZ is dynamic process of filaments assembling/disassembling, the disruption of polymerization as well as depolymerization leads to disorder of the cell cycle. Thus, we have identified several benzimidazole derivatives with big potential for further development of new antibacterial drugs.

Acknowledgments. The research was supported by STCU project #4932 "Design of new benzimidazole and phenylcarbamate compounds with increased activity against mycobacterial FtsZ"

1. E.D. Brown, G.D. Wright (2005) New targets and screening approaches in antimicrobial drug discovery, *Chem. Rev*., **105**: 759–774.
2. W.Vollmer (2006) The prokaryotic cytoskeleton: a putative target for inhibitors and antibiotics?, *Appl. Microbiol. Biotechnol*., **73**(1): 37-47.
3. M. Sarcina, C.W. Mullineaux (2000) Effects of tubulin assembly inhibitors on cell division in prokaryotes in vivo, *FEMS Microbiol Lett.*, **191**(1): 25-29.
4. R.A. Slayden et al. (2006) Identification of cell cycle regulators in Mycobacterium tuberculosis by inhibition of septum formation and global transcriptional analysis, *Microbiology*, **152**: 1789-1797.
5. K. Kumar et al. (2011) Novel trisubstituted benzimidazoles, targeting Mtb ftsZ, as a new class of antitubercular agents, *J. Med. Chem.*, **54**: 374–381.
6. D.W. Ritchie, V. Venkatraman (2010) Fast FFT protein-protein docking on graphics processors, *Bioinformatics*, **26**: 2398-2405.
7. D.J. Haydon et al. (2010) Creating an antibacterial with in vivo efficacy: synthesis and characterization of potent inhibitors of the bacterial cell division protein ftsZ with improved pharmaceutical properties, *J. Med. Chem.*, **53**(10): 3927-3936.
8. J.M. Andreu et al. (2010) The antibacterial cell division inhibitor PC190723 is an ftsZ polymer-stabilizing agent that induces filament assembly and condensation, *J. Biol. Chem.*, **285**(19): 14239-14246.

Systems Modeling of EphB4/ephrinB2 Signaling Pathways

Artem Demidenko[1], Kirill Peskov[1], Aleksandr Dorodnov[1], Oleg Demin[1],
Kenneth Luu[2], Eugenia Kraynov[2], Dawn Nowlin[2]

[1]*Institute for Systems Biology SPb, Moscow, Russia, Russian Federation, artemdem@insysbio.ru*
[2]*Pfizer Global Research and Development, La Jolla, California, USA, United States*

It has recently been shown that Eph-ephrin interactions play an essential role not only in tumor angiogenesis but also in tumor progression and/or suppression. The exact mechanism contributing to such a multitude of responses, however, remains unclear. In order to better understand this problem we studied the intricacies of EphB4 biology using a systems modeling approach. The main aims of this study were to: (1) reconstruct the EphB4/ephrinB2 signaling pathways based on information mined from the literature; (2) develop a kinetic model for EphB4/ephrinB2 forward signaling, and (3) analyze the model behavior and prediction to gain deeper insight into EphB4/ephrinB2 forward signaling and its influence on tumor progression and/or suppression mechanisms.

DBSolve Optimum software was used for all model development and analysis steps. Kinetic model verification was performed using data from ephrinB2-Fc induced EphB4 activation, internalization and degradation obtained in a MCF7 breast cancer cell line and other datasets (approx. 30 datasets) published in the literature.

The signaling pathways of EphB4-ephrinB2 interactions were successfully reconstructed and a kinetic model of EphB4-ephrinB2 forward signaling was developed. Analysis of the model behavior allowed us to make the following predictions about the system regulation and possible cellular responses at different physiological conditions: (1) the Abl/Rac/Rap branch of EphB4-induced forward signaling was cell specific, (2) one of the main reasons for (1) was an inhibition of Abl function by filament actin. In addition, it followed that cell types with high concentration of filament actin were less sensitive to EphB4 inhibition of migration, (3) cell proliferation potential had low sensitivity against EphB4 activation and (4) EphB4 activation had an important influence on AKT phosphorylation.

It was shown that EphB4-ephrinB2 forward signaling had a negative effect on cell proliferative, survival and migratory potential. However, those effects were cell specific and can be weakened in certain cell types. In these cases, increasing the concentration of filament actin could completely inhibit the Rac/Rap branch of EphB4-ephrinB2 signaling.

Human-chimpanzee property-dependant conservation

Igor Deyneko, Helmut Bloker

Department of Genome Analysis, HZI, Braunschweig, Germany, ide@helmholtz-hzi.de

We performed a comparative analysis of promoters of orthologous genes of human and chimpanzee located on chromosomes 21. Similarities between DNA sequences were calculated using property-dependent similarity measure [1] (using FeatureScan [2] with "melting enthalpy") and letter-based similarity (using ClustalW).

The distribution of the number of promoters vs. similarity reveals that promoters of chimpanzee and human genes, which differ by less than 2% of nucleotides, show significantly higher similarity by FeatureScan than expected (under pure random and transition/transversion biased models). We found that 139 out of 198 orthologous promoter pairs showed higher signal similarity than can be expected - with a p-value of $2{,}43*10^{-8}$.

Using the EMBL-EBI gene ontology classification, we examined the distribution of genes, which showed high signal similarity of their promoters. A subset of 15 genes involved in the molecular function "metal ion binding" and another subset of 11 genes involved in "nucleotide binding" with p-values of $4{,}9*10^{-3}$ and $4{,}52*10^{-5}$, respectively. Contrarily, applying ClustalW, no significant association with GO terms was found.

A strong correlation of signal similarity with gene expression was found by comparing our present results with our earlier data [3]. The prediction accuracy of gene expression by comparing their promoter sequences reaches a sensitivity of 83% and a selectivity of 60% using FeatureScan. At the same time, using letter conservation under the same conditions provides only 30% and 40%, respectively.

As we may conclude from the presented results, a nucleotide substitution decreasing the melting temperature of a locus, may induce evolutionary pressure for further changes in the close vicinity to compensate the first mutation. As further investigations, it might be interesting to investigate changes of characteristics (melting temperature, conformation and others) caused by SNP mutations and its correlations with phenotype or diseases.

1. Kauer G. et.al., (2003) Applying signal theory to the analysis of biomolecules. *Bioinformatics*, **19**, 2016-2021.
2. Deyneko IV. et.al., (2006) FeatureScan: revealing property-dependent similarity of nucleotide sequences. *Nucleic Acids Research*, **34**, W591-W595.
3. Watanabe, H. et al., (2004) DNA sequence and comparative analysis of chimpanzee chromosome 22. *Nature*, **429**, 382-388.

Translational Studies in the Genomic Era

Luda Diatchenko

University of North Carolina at Chapel Hill, United States, lbdiatch@email.unc.edu

A presentation will present an illustration of how human association study results can then be translated into the pharmacological treatment of common clinical conditions, exemplified by genetic variants of the catechol-O-methyltransferase (COMT) gene, an enzyme that metabolizes catecholamines. First, three major haplotypes of COMT, designated as low pain sensitive (LPS), average pain sensitive (APS), and high pain sensitive (HPS) have been identified based on a carrier's response to experimental pain stimuli. These three haplotypes account for 11% of the variability to human experimental pain sensitivity and are predictive of the risk of onset of a common musculoskeletal pain disorder (i.e., temporomandibular disorders, or TMD). Next, we showed the molecular genetic mechanism thereby the LPS haplotype produces higher levels of COMT enzymatic activity than the APS or HPS haplotypes. Third, we demonstrated that the pharmacological inhibition of COMT in rats results in mechanical and thermal hypersensitivity that is reversed by the nonselective β-adrenergic antagonist propranolol, or by the combined administration of selective β2- and β3-adrenergic antagonists. These data provide the first direct evidence that low COMT activity leads to increased pain sensitivity via a β2/3-adrenergic mechanism, and suggests that pain conditions associated with low COMT activity and/or elevated catecholamine levels can be treated with non-selective β-blockers. This finding led to the clinical studies showed that propranolol, a clinically used non-selective β-adrenergic antagonist which is widely used for treatment of hypertension, may be and effective treatment for chronic pain conditions in a manner that is dependent on the subject's COMT diplotype. Thus, COMT haplotypes may serve as genetic predictors of treatment outcomes and permit the identification of a subgroup of patients who will benefit from propranolol therapy.

Comparative analysis of lipid biosynthesis in Archaea and Bacteria: What was the structure of first membrane lipids?

Daria Dibrova[1,2], Kira Makarova[3], Michael Galperin[3],
Eugene Koonin[3], Armen Mulkidjanian[1,4]

[1] *School of Physics, University of Osnabrück, D-49069 Osnabrück, Germany, udavdasha@belozersky.msu.ru*
[2]*School of Bioengineering and Bioinformatics, Moscow State University, Moscow 119992, Russia*
[3]*National Center for Biotechnology Information, NLM, NIH, Bethesda, MD 20894, USA,*
[4]*A.N. Belozersky Institute of Physico-Chemical Biology, Moscow State University, Moscow 119899, Russia*

The chemical nature of the evolutionarily first membranes, which should have encased the Last Universal Common Ancestor (LUCA), cannot be inferred easily because the chemical compositions and biogenesis pathways of archaeal and bacterial phospholipids are fundamentally different [1-3]. However, the nearly universal conservation of the elements of general protein secretory pathway [4] and of the F- and V-type ATP synthases [5, 6], which are membrane-embedded molecular machines, suggests that the LUCA did possess some kind of membranes [7-9].

Biosynthesis of membrane lipids, both in bacteria and in archaea, proceeds in three major steps, see [2, 3] for reviews. First, either fatty acids (in bacteria) or isoprenoid chains (in archaea) are synthesized. While there is only one synthetic pathway for fatty acids, there are two synthetic pathways for the 5-carbon units of the isoprenoid chains, namely the mevalonate pathway [10] and the deoxyxylulose phosphate (DXP) pathway [11]. Next, the hydrophobic tails (synthesized at the first step) are connected by ester (bacteria) or ether (archaea) bonds, respectively, to the specific stereoisomers of glycerol phosphate (sn-glycerol-3-phosphate (G3P) in bacteria and sn-glycerol-1-phosphate (G1P) in archaea). Lastly, diverse polar head groups are attached to the glycerol moieties. The difference between archaeal and bacterial membranes extends beyond the chemical structures of the phospholipids, to the evolutionary provenance of the enzymes involved in membrane biogenesis that are either non-homologous or distantly related but not orthologous in bacteria and archaea [2, 3].

Using the available complete genomes, we have analyzed the occurrence of enzymes involved in these three major reactions of lipid biosynthesis in the bacterial and archaeal phyla, relying on the assignment of respective enzymes to the Clusters of Orthologous Clusters of Orthologous Groups of proteins (COGs) [12]. Our analysis indicates, in agreement with [10], that the mevalonate pathway of isoprenoid biosynthesis preceded the DXP pathway in evolution and may have already operated in the LUCA. In contrast, the enzymes responsible for the

synthesis of fatty acids and their attachment to the G3P moiety are found only within bacteria, so that the fatty-acid-based phospholipids should have been a bacterial invention. Based on our analysis, we hypothesize that (i) either the enzyme for the G1P synthesis were present in the LUCA, so that the a G1P unit was a constituent of the first isoprenoid phospholipids, or (ii) the glycerol moiety was absent from the lipids of the LUCA, as suggested earlier [13].

1. Smit A, Mushegian A (2000) Biosynthesis of isoprenoids via mevalonate in Archaea: the lost pathway. *Genome Res*, **10**(10):1468-1484.
2. Pereto J, Lopez-Garcia P, Moreira D (2004) Ancestral lipid biosynthesis and early membrane evolution. *Trends Biochem Sci*, **29**(9):469-477.
3. Koonin EV, Martin W (2005) On the origin of genomes and cells within inorganic compartments. *Trends Genet*, **21**(12):647-654.
4. Cao TB, Saier MH, Jr. (2003) The general protein secretory pathway: phylogenetic analyses leading to evolutionary conclusions. *Biochim Biophys Acta*, **1609**(1):115-125.
5. Mulkidjanian AY, Makarova KS, Galperin MY, Koonin EV (2007) Inventing the dynamo machine: the evolution of the F-type and V-type ATPases. *Nat Rev Microbiol*, **5**(11):892-899.
6. Dibrova DV, Galperin MY, Mulkidjanian AY (2010) Characterization of the N-ATPase, a distinct, laterally transferred Na+-translocating form of the bacterial F-type membrane ATPase. *Bioinformatics*, **26**(12):1473-1476.
7. Jekely G (2006) Did the last common ancestor have a biological membrane? *Biol Direct*, **1**:35.
8. Mulkidjanian AY, Galperin MY, Makarova KS, Wolf YI, Koonin EV (2008) Evolutionary primacy of sodium bioenergetics. *Biol Direct*, **3**:13.
9. Mulkidjanian AY, Galperin MY, Koonin EV (2009) Co-evolution of primordial membranes and membrane proteins. *Trends Biochem Sci*, **34**(4):206-215.
10. Lombard J, Moreira D (2011) Origins and early evolution of the mevalonate pathway of isoprenoid biosynthesis in the three domains of life. *Mol Biol Evol* , **28**(1):87-99.
11. Eisenreich W, Bacher A, Arigoni D, Rohdich F (2004) Biosynthesis of isoprenoids via the non-mevalonate pathway. *Cell Mol Life Sci*, **61**(12):1401-1426.
12. Tatusov RL, Galperin MY, Natale DA, Koonin EV (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res*, **28**(1):33-36.
13. Mulkidjanian AY, Galperin MY (2010) Evolutionary origins of membrane proteins In: *Structural Bioinformatics of Membrane Proteins*. Edited by Frishman D. Viena: Spriger; 1-28.

Practicality and Time Complexity of a Sparsified RNA Folding Algorithm

Slavica Dimitrieva, Philipp Bucher

Swiss Institute for Experimental Cancer Research (ISREC), School of Life Sciences, Swiss Federal Institute of Technology (EPFL), CH-1015 Lausanne, Switzerland, slavica.dimitrieva@epfl.ch, philipp.bucher@epfl.ch

Commonly used RNA folding programs compute the minimum free energy structure of a sequence under the pseudoknot exclusion constraint. They are based-on Zuker's algorithm [1] which runs in time $O(N^3)$. Recently, it has been claimed that RNA folding can be achieved in time $O(N^2)$ using a sparsification technique that exploits the "triangular inequality" [2]. Several variants of sparse RNA folding were proposed [3]. Here, we present our own version, which is readily applicable to existing RNA programs as it does not require any new data structure. Speed-gain is achieved solely by re-ordering and conditional execution of elementary arithmetic operations. The principle is explained in Fig. 1 using the simpler Nussinov algorithm, which maximizes the number of base-pairs. The same trick is applicable to energy-minimization based RNA folding and a variety of related algorithms.

a)
for $i = N$ *to* 1
  for $j = i+1$ *to* N {
    $F(i,j) = \max\{F(i+1,j-1)+\delta(i,j), F(i,j-1), F(i+1,j)\}$
    for $k = i+1$ *to* $j-1$
      $F(i,j) = \max\{F(i,j), F(i,k)+F(k+1,j)\}$
  }

b)
for $j = 1$ *to* N
  for $i = j$ *to* 1 {
    $F(i,j) = \max\{F(i,j), F(i,j-1), F(i+1,j)\}$
    if $(F(i+1,j-1)+\delta(i,j) > F(i,j))$ {
      $F(i,j) = F(i+1,j-1)+\delta(i,j)$
      for $k = i-2$ *to* 1: $F(k,j) = \max\{F(k,j), F(k,i-1)+F(i,j)\}$
    }

Figure 1. Pseudo-code for the original Nussinov algorithm (a) and our sparsified version (b). *F(i,j)* contains the best score, i.e. the maximal number of bases pairs, for the subsequence starting at position *i* and ending at position *j*. The function *δ(i,j)* returns 1 if the bases at positions *i* and *j* can form a canonical base pair, and 0 otherwise.

To prove practicality, we applied the proposed modification to the widely used Vienna RNAfold program, to create sibRNAfold, the first publicly available sparsified RNA folding program. We deliberately opted for modification of existing code rather than reprogramming to ensure identical results with a trusted implementation. Using real RNA sequences, we observed up to 50-fold speed gain for long RNAs. The time for folding the HIV genome (~10 kb) went down from 344 sec to 19. The SARS (~30 kb) genome was folded in 6 min compared to nearly 3 hrs required by Vienna RNAfold. Using sibRNAfold, we were able to fold the 100kb long Titin mRNA in only 3 hrs.

To gain a better understanding of the time complexity of sparsified RNA folding in general, we carried out a thorough run time analysis with synthetic random sequences, both in the context of energy minimization and base-pairing maximization. Contrary to previous claims, the asymptotic time complexity of sibRNAfold appears to be $O(N^3)$ under a wide variety of conditions (Fig 2a,b). Specifically, we varied the base composition, the folding temperature, and multi-loop parameters of the energy function. A previous proof of quadratic time complexity was based on the assumption that computational RNA folding obeys the "polymer zeta property" [2]. Specifically, the polymer zeta property stipulates that the probability that the terminal bases of a folded RNA are paired, exponentially decreases to zero with increasing sequence length. Consistent with our run-time analysis, we found that RNA folding does not obey the polymer zeta property. Different and more interesting results were obtained with a sparsified Nussinov algorithm. While run-time remained cubic for RNA sequences with an unbiased base composition, we observed quadratic behavior with sequences enriched in A and C. There appears to be a sharp phase transition at about 57% A+C content both with regard to time complexity (Fig 2c) and polymer zeta property. The code used in this work is available at: http://sibRNAfold.sourceforge.net/

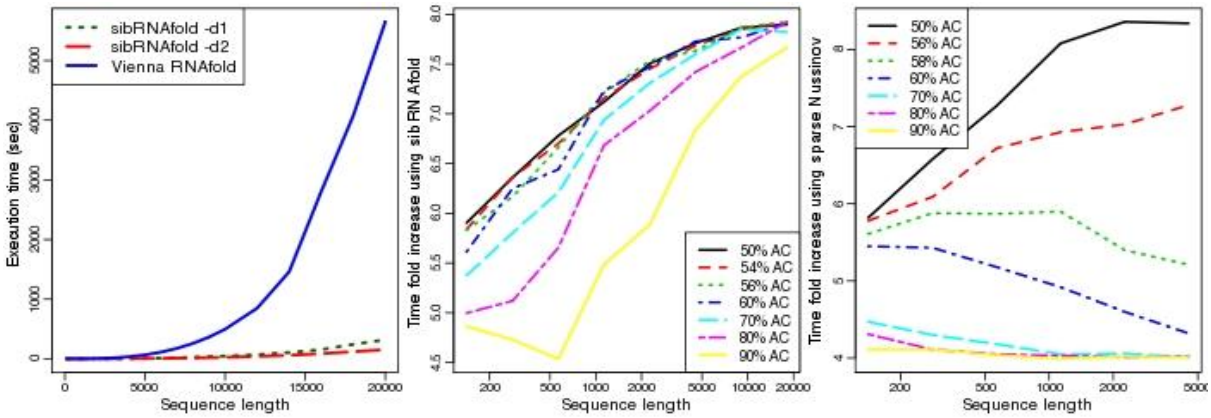


Figure 2. a) Speed comparison of sibRNAfold vs Vienna RNAfold program; b,c) Sequence base composition dependence of the speed gain using sibRNAfold (b) and using sparsified Nussinov algorithm (c). The vertical axis (time fold increase) reflects the fold increase in the number of multiloop operations resulting from doubling the sequence length. All the values represent an average over 100 random sequences the same length.

1. M. Zuker, P. Stiegler. (1981), *Nucl. Acids Res*, **9:** 133–148.
2. Y. Wexler et al. (2007), *J. Computat. Biol.*, **14**: 856-72
3. R. Backofen et al. (2011) *JDA* 9:12-13.

A Quantitative Systems Pharmacology Model Provides Insights into Phosphate Homoeostasis through Multiple Interacting Pathways

ALEKSANDR DORODNOV[1], KIRILL PESKOV[1], ARTEM DEMIDENKO[1], OLEG DEMIN[1], BALAJI AGORAM[2]

[1]*Institute For Systems Biology SPb Moscow, Russian Federation, dorodnov@insysbio.ru*
[2]*MedImmune UK, United Kingdom*

Phosphate is an important mineral required for numerous cellular functions such as DNA and membrane lipid synthesis, generation of high energy phosphate esters, and intracellular signalling. However, an integrated understanding of phosphate regulation, the various control mechanisms, and interactions between the mechanisms is not available yet. We describe here the first attempt, to our knowledge, to develop an integrated quantitative understanding of the factors responsible for endogenous phosphate regulation. Based on Bergwitz (2010), a minimal model consisting of bone, serum, parathyroids, intestine, and kidney tissues and fibroblast growth factor (FGF) 23, parathyroid hormone (PTH), VitaminD, and phosphate entities was developed. The known dependencies (Ben-Gov 2007) were hypothesised based on literature and quantitatively characterised based on available data. The individual submodels were integrated to provide a unified model of phosphate homoeostasis. Model predictions were verified with available literature data on therapies known to impact phosphate levels (e.g. FGF23, and FGFR modulators) and from human genetic disorders. We then used the model to simulate single and multiple dose phosphate changes for different FGFR modulators that can potentially alter endogenous phosphate levels. All individual submodels provided adequate descriptions of isolated interactions. The magnitude, but not time course of Vit D changes after FGFR modulation was predicted correctly. The integrated model provided good descriptions of literature-reported changes in phosphate, and Vit D levels after FGFR therapy. FGFR-subtype specific modulators had generally different VitD and phosphate effects. The contribution of PTH to overall phosphate homoeostasis was found to be not significant compared to that of bone (through FGF23). A minimal QSP model of phosphate homoeostasis was developed. The model can be used to evaluate the potential effect of various therapeutic options affecting the phosphate homoeostasis pathway.

1. Bergwitz and Juppner (2010) Regulation of phosphate homeostasis by PTH, vitamin D, and FGF23, Annu Rev Med, 61:91-104.
2. Ben-Gov (2007), J Clin Inv.

Protein-Membrane Binding as a Self-Adapting Process: a Computational View

Anastasia Konshina, Darya Pyrkova, Anton Polyansky, Roman Efremov

M.M. Shemyakin & Yu.A. Ovchinnikov Institute of Bioorganic Chemistry, Russian Academy of Sciences, Russian Federation, efremov@nmr.ru

Cell membranes, along with their individual components like membrane-bound proteins, particular lipids or lipid bilayer itself, attract a growing attention as very perspective pharmacological targets. According to recent estimations, up to 70% of currently marketed drugs act on these targets. Rational design of new highly efficient and selective compounds (drug prototypes) modulating activity of biomembranes, requires atomic-scale information on their spatial structure and dynamics under different conditions. Because such details resist easy experimental characterization, important insight can be gained via computer simulations.

Here, we present the results of structural/dynamic studies of two classes of membrane active agents - cationic peptides with antimicrobial activity (AMPs) [1] and cardiotoxins from snake venom (CTXs), which are capable of lysing red blood cells [2]. The computational approach combines in a self-consistent manner Monte Carlo conformational search in implicit hydrophobic slabs and molecular dynamics in hydrated full-atom lipid bilayers composed of different types of lipids. Despite different structure (AMPs are α-helical, while CTXs have β-structural fold) and mechanism of membrane permeation (AMPs exhibit non-specific membrane binding, while CTXs may also specifically interact with anionic lipids in bilayers), in both cases the process of polypeptide interaction with cell membranes reveals a prominent “self-adapting” character. Namely, both classes of membrane active agents employ a wide arsenal of structural/dynamic tools in order to accomplish their function – lysis of cell membrane. Importantly, the lipid bilayer of biological membranes plays an essential role in the recognition and binding events. In particular, the presence of anionic lipids induces formation of highly dynamic lateral heterogeneities (clusters) on the membrane surface, which differ in their packing and hydrophobic properties from the bulk lipids. Such a mosaic nature of membranes is tuned in a wide range by the chemical nature and relative content of lipids, presence of ions, and so on [3]. This makes possible mutual adaptation of the two amphiphatic systems (peptide and membrane). In our opinion, such a diversity of the factors important for polepeptide-bilayer interactions assures their efficient and robust binding to cell membranes. Understanding of such effects creates a basis for rational design of new physiologically active molecules and/or artificial membranes with predefined properties.

This work was supported by the Russian Foundation for Basic Research, by the RAS Programme “Basic fundamental research of nanotechnologies and nanomaterials”, and by the Ministry of Science and Education of the Russian Federation. Access to computational facilities of the Joint Supercomputer Center RAS (Moscow) and Computer Center of M.V. Lomonosov Moscow State University is gratefully acknowledged.

1. A.A. Polyansky, R.Ramaswamy, P.E. Volynsky, I.F. Sbalzarini, S.J. Marrink, R.G. Efremov (2010) Antimicrobial peptides induce the growth of nascent phosphatidylglycerol domains in a model bacterial membrane. J. Phys. Chem. Lett. 1:3108-3111.
2. A.G. Konshina, I.A. Boldyrev, Yu.N. Utkin, A.V. Omel’kov, R.G. Efremov (2011) Snake cytotoxins bind to membranes via interactions with phosphatidylserine head groups of lipids. PLoS ONE 6:e19064.
3. D.V. Pyrkova, N.K. Tarasova, T.V. Pyrkov, N.A. Krylov, R.G. Efremov (2011) Atomic-scale lateral heterogeneity and dynamics of two-component lipid bilayers composed of saturated and unsaturated phosphatidylcholines. Soft Matter 7:2569-2579.

Microevolutionary changes in conditions of chronic environmental stress

Mariya Elkina, Darya Pyrkova, Tatyana Glazko

Russian State Agrarian University - MTAA, Moscow, 127550, Timirjazevsky street, 49, *vglazko@yahoo.com*

Influence of ecological conditions on the genetic structure of populations is one of the leading factors of intra- and interspecies differentiation. However, the molecular-genetic mechanisms of such processes are still insufficiently investigated. In order to study the species specificity of the responses to the chronic effect of environmental stress factors a comparative analysis of population-genetic structure of the Mongolian livestock (yaks, cattle, sheep) in a Biosphere Reserve Houbsougul (Mongolia) and in the zone of risky animal breeding Gobi Desert was carried out. The estimations of polymorphism of DNA fragments flanked by inverted microsatellite repeats (AG, ISSR-PCR markers) were used as molecular genetic markers. In order to evaluate the influence of different ecological conditions of livestock on the stability of the genetic apparatus in smears of peripheral blood of animals the counting of erythrocytes with micronuclei was performed. Genotyping of animals for 38 loci yielded the following data. The rate of polymorphic loci (P) of the yak was 15,8%, polymorphic information content (PIC) 0,039. In the populations of yak of the Gobi Desert – P - 10,5%, PIC - 0,039; of the reserve Houbsougul - P 13,2%, PIC - 0,040, the genetic distance between them DN = 0,0410. In Mongolian cattle it was revealed in total P - 28,9%, PIC - 0,049; in the Gobi Desert – P - 13,2%,

PIC - 0,035; in the reserve Houbsougul – P - 23,7%, PIC - 0,067; genetic distance between them DN = 0.1087. In Mongolian sheep it was observed in total P - 42,1%, PIC - 0,095, in the Gobi Desert – P - 34,2%, PIC - 0,118; in the reserve Houbsougul – P - 10,5%, PIC - 0,071; genetic distance between them DN = 0,1712. On the dendrogram constructed on the genetic distance values, yaks form a common cluster with cattle, both groups of sheep - a single cluster. The obtained data indicated that the population-genetic differentiation on the ISSR-PCR markers was the most profound in sheep, the least - in yak, which was consistent with the level of polymorphism in these species (P, PIC). The frequency of erythrocytes with micronuclei in the area of Houbsougul was significantly higher than that of the same species in zone of the high-risk animal breeding. The frequency of erythrocytes with micronuclei in the area of Houbsougul in sheep was 5,3 ± 0,4 ‰; in cattle - 4,6 ± 0,7 ‰; in yak - 3,2 ± 0,6 ‰; in the Gobi Desert in sheep - 0,9 ± 0,1 ‰; in cattle - 1,8 ± 0,6 ‰; in yak - 0,3 ± 0,2 ‰. That is, the differentiation of animals on the micronuclei test coincided with that observed in molecular-genetic markers: the greatest - in sheep, the lowest - in yak. The frequency of occurrence of red blood cells with micronuclei in cattle and sheep of Houbsougul corresponds to typical for a number of other breeds of these species; we investigated earlier, in contrast to the animals of the Gobi Desert. Micronuclei test in somatic cells is widely used for bioindication of ecotoxical effects as an indicator of instability of the genetic apparatus. It is known that the frequency of cells with cytogenetic abnormalities in peripheral blood cells in mammals is closely linked with reproductive dysfunction (Rubes et al., 1991; Migliore et al., 2006). The data obtained suggest that all three species in the Gobi Desert were most resistant to the adverse effects in relation of animal selection (in several generations), with increased stability of the genetic apparatus. We also found that interspecies genetic distances for each species were less than between groups of animals Houbsougul and each pair of groups of two other species. For example, between the yaks of the Gobi Desert and both groups of the cattle DN – 0,3318, 0,3758, and between the yaks of Houbsougul and these same groups of the cattle – 0,2759, 0,2528. Apparently as a result of natural selection a unique gene pool was formed in the zone of risky animal breeding, the specificity of which can be identified using a random sample of ISSR-PCR markers.

1. Rubes J., Horinova Z., Gustavson I., Borkovec L., Urbanova J. (1991) Somatic chromosome mutations and morphological abnormalities in sperms of boars, *Hereditas*, **115:** 139-143.
2. Migliore L., Colognato R., Naccarati A, Bergamaschi E. (2006) Relationship between genotoxicity biomarkers in somatic and germ cells: findings from a biomonitoring study, *Mutagenesis*, **21 (2):** 149-152.

The genetic structure of musk oxen populations, using ISSR-PCR markers

Irina Elsukova, Taras Sipko, Nikolay Badrukov, Valery Glazko

Russian State Agriculture University – MAA named after K.A. Timiryazev, Russian Federation,
irinaelsukova@gmail.com

Population decrease of several species of ungulates leads to necessity for their condition control and for work on their reintroduction. In this regard, three population's of musk oxen (inhabiting on the Taimyr Peninsula, on the Wrangell Island and on the Greenland Island) genetic structure was studied using ISSR-PCR (Inter-Simple Sequence Repeat). Twenty animals were brought to the Wrangell Island from the Nunivak Island (the USA). From the same initial population on the Taimyr Peninsula 20 musk oxen and 10 animals from the Banks Island (Canada) were inhabited.

To the Nunivak Island the population have been imported from eastern Greenland. Population of musk oxen inhabited at the Banks Island was sufficiently large and heterogeneous. West Greenland population (the Greenland Island) in our studies was used for comparison as successfully reintroduced (musk-ox-founders were moved from the native populations of the eastern part of Greenland). To investigate the genetic structure of populations (estimate of the proportion of polymorphic loci, polymorphic information content of each locus - PIC) ISSR-PCR method was used. As the primers in polymerase chain reaction (PCR) fragments of di-and trinucleotide microsatellites with anchor nucleotides (AG)9C, (GA)9C, (GAG)6C was used.

Each amplification product was considered as a separate locus.

Total polymorphism was estimated by 20 loci. Monomorphic (conservative) and polymorphic DNA fragments of specific length, flanked by inverted repeats of these primers was defined.

Monomorphic for all populations of musk oxen were the following loci: in the amplicons spectra of PCR primer (AG)9C DNA fragment length 450 bp, primer (GAG) 6C - 600 and 510 bp, primer (GA) 9C - 570 and 500 bp. In the spectra of PCR primers (GAG) 6C only among musk oxen of the Wrangell Island's population a DNA fragment 910bp in lengt was met, this fragment apparently can be a marker for this population. Comparative analysis of PIC values for all three populations of musk oxen indicates that the population of the Wrangell Island is different from the other two with the lowest average heterozygosity. Genetic distances, reflecting the specific characteristics of genetic differentiation between the studied populations, were

highest between a group of animals from the Wrangell Island and the other two, in comparison with the genetic distance between populations of the Greenland Island and Taimyr Peninsula.

On the basis of calculating the values of genetic distances by the method of Nei, 1972 constructed a general dendrogram that takes into account all the received loci showing genetic relationships between the studied groups of animals.

Obtained data shows the complexity of the genetic processes occurring in generations reintroduced on the Wrangell Island and the Taimyr Peninsula populations of musk oxen.

Spectrum of the amplification products can identify a number of molecular genetic markers of various genomic sites, which may be used later to control the dynamics of the genetic structure of each population and to take appropriate measures to preserve population. The lowest level of heterozygosity (PIC) we have identified at musk ox of Wrangell's population, which may indicates its determine a disadvantage, probably due to the relatively high level of inbreeding. Comparing information about importation of animals and the dates of our study, it must be concluded that for introduction as a founding population the maximum possible number of animals from heterogenous population or from genetically unrelated population should be used.

CodY regulon in Bacillaceae

Ekaterina Ermakova[1], Mikhail Gelfand[1], Dmitry Rodionov[1,2]

[1]*Institute for information transmission problems RAS, Russian Federation, ermakova@iitp.ru*
[2]*Burnham Institute for Medical Research, United States*

Transcription factor CodY is a global regulator of nutrient limitation and amino acid metabolism in Firmicutes studied mostly in B. subtilis. Starting with a set of 42 experimentally verified binding sites in B. subtilis (Belitsky and Sonenshein 2008 and B. Belitsky, personal communication), CodY regulon was analysed in Bacillaceae using comparative genomics methods. We show that CodY regulon in Bacillaceae consists of at least 135 clusters of orthologous operons having conserved binding sites.

Diversity of the restriction-modification systems in full prokaryotic genomes

Anna ERSHOVA[1,3], Sergei SPIRIN[1,2], Anna KARYAGINA[1,3], Andrei ALEXEEVSKI[1,2]

[1]*Belozersky Institute for Physical and Chemical Biology, Moscow State University*
[2]*Scientific Research Institute for System Studies (NIISI RAS), Moscow*
[3]*Gamaleya Institute of Epidemiology and Microbiology, Moscow, asershova@gmail.com*

Restriction-modification (R-M) systems prevent host cells from invasions of foreign (e.g., phages) DNA. A typical R-M system has two activities: (1) DNA cleavage in specific sites, which results in destroying of the foreign DNA; (2) DNA methylation of the same sites, preventing cleavage of the host DNA. In addition to anti-phage protective role of R-M systems, they were hypothesized to be selfish genetic elements [1]. Each activity of an R-M system may correspond to a single protein (a restriction endonuclease (RE), a DNA-methyltransferase (M), or a fusion RM protein), or a complex of proteins. For all annotated R-M systems, genes of separate RE and M proteins are linked in prokaryotic genome [2, 3]. R-M systems are classified into four types (I–IV) [2]. In this work we deal with R-M systems of types I, II, and III, which cleave non-modified DNA sites.

To reveal diversity of R-M systems, we have studied the distribution of R-M systems (data from REBASE [3]) in 1032 complete prokaryotic genomes (including available data on plasmids). Here we discuss: (i) life-style of the organisms free of R-M systems and ones possessing extremely high number of them, (ii) distribution of complete and incomplete (i.e. lacking RE or M gene) R-M systems, and (iii) existence of a number of scattered R-M systems.

We found 75 genomes free of R-M systems, 73% of them are genomes of intracellular endosymbionts. As much as 65% of intracellular endosymbionts and only 2% of free living prokaryotes are free of R-M systems. Absence of R-M systems highly correlates with absence of CRISPR cassettes, another prokaryotic anti-phages protection system. Data on distribution of CRISPR cassettes in complete prokaryotic genomes were downloaded from IMG database [4]. We suppose that most prokaryotes having no known protective systems actually do not meet any bacteriophages in their life. These data in association with Ichizo Kobayashi hypothesis [1] allows to interpret the relations between R-M systems and their hosts as mutualistic.

In 957 genomes we identified 2257 complete and 3017 incomplete R-M systems. The majority of prokaryotic genomes carry 1–4 complete and 1–4 incomplete R-M systems. Only a small fraction of incomplete R-M systems have known biological role (Dam methyltransferases are among them). Thus, either genes of incomplete R-M systems play yet unknown biological role or incomplete R-M systems are disabled. In the latter case, a large number of incomplete R-M systems points to active R-M expanding that confirms Kobayashi hypothesis.

Among incomplete R-M systems we found 135 putatively active RE genes lacking paired active M gene in its vicinity (solitary RE genes). Among them there are 62 of type I, 38 of type II and 35 of type III. Note that bacteria with a functional RE gene of type II R-M systems cannot survive without corresponding active M gene (single RE molecule of such systems will cleave the unmodified host DNA). We proposed that in such cases the cell death may be prevented by a gene of a paired active M localized rather far from the RE gene in the genome. Here such R-M systems are called scattered R-M systems.

For each solitary RE gene and a candidate to a paired M gene, genomes carrying orthologs of both RE and M genes were determined. Bidirectional Best Hits method was used for orthologs detection. Such pairs (RE gene, M gene) were considered as putative scattered R-M systems. 19 scattered R-M systems of type I and 10 scattered R-M systems of type II were found (see Fig. 1 for examples). Existence of remaining solitary RE genes may be explained either by the lost of their functionality or by a presence of a paired M gene in a plasmid genome.

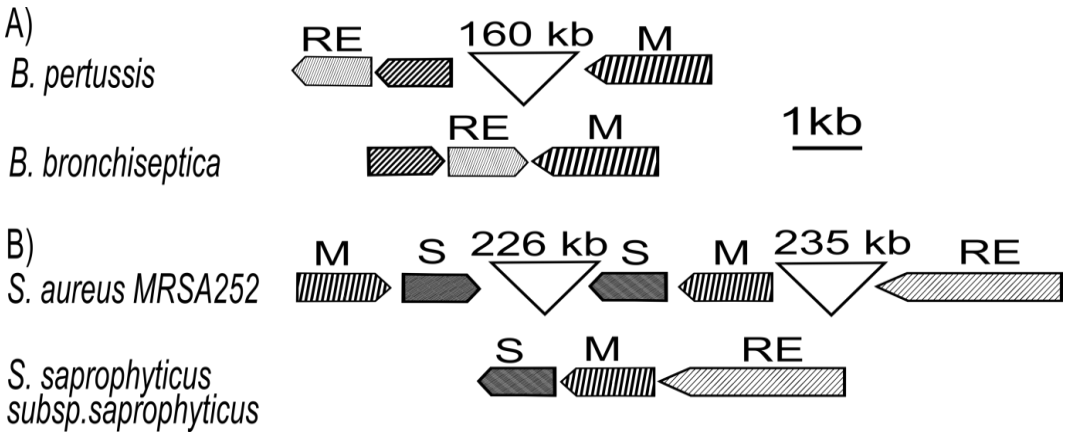


Fig. 1. Scattered R-M systems. Orthologous proteins are filled similarly. A) Scattered R-M system of type II in the genome of Bordetella pertussis compared to the R-M system of Bordetella bronchiseptica. B) Scattered R-M system of type I in the genome of Staphylococcus aureus MRSA252 compared to the R-M system of Staphylococcus saprophyticus (see text for explanation).

In [5] activity of genes of RE, and both paires of M and S subunits (S is the recognition subunit for type I R-M system) of *S. aureus* (Fig.1B) was demonstrated experimentally. Functional cleavage complex RE_2M_2S can be formed either by one pair of (M,S) genes, or by another pair. On the best of our knowledge, this is the only described in literature example of R-M system, considered as scattered in our terminology.

Our findings of other scattered systems, in the case of experimental confirmations, may shed light on complicated relations between RE and M genes in host genomes.

1. Kobayashi I. 2001. *Nucl. Acids Res.* **29**: 3742–3756.
2. Wilson G.G. 1991. *Nucl. Acids Res.* **19:** 2539–2566.
3. Roberts, R.J., Vincze, T., Posfai, J., Macelis, D. 2010. *Nucl. Acids Res.* **38**: D234-D236.
4. Kyrpides N. C. et al. 2010. *Nucl. Acids Res.* **38:** D382-D390.
5. Lindsay J. A., Waldron D. E. 2006. *Journal of Bactriology.* **188** (15): 5578–5585.

Detecting Selection via Model-based Geographical Mapping

Wen-Yun Yang[1], Eleazar Eskin[2], Eran Halperin[3]

[1]*University of California, Los Angeles, United States, wenyun.yang@gmail.com*
[2]*University of California, Los Angeles, United States, eeskin@cs.ucla.edu*
[3]*Tel Aviv University, Israel, heran@post.tau.ac.il*

In the last two decades, analysis of genetic variation data, particulalry single nucleotide polymorphisms (SNPs), has been widely applied for the study of populations, and resulted in the discovery of genetic risks for disease, as well as the characterization of evolutionary processes, such as the inference of recombination hotspots, and regions under selection. The characterization of the genetic diversity across different populations has been playing a central role in these studies; particularly, such characterization is used for population stratification in genome-wide association studies [1], and for the discovery of novel associations of SNPs to disease through admixture mapping. Moreover, by exploiting the differnces in genetic variation across different populations, it has been possible to detect genes that have been under selection, and to learn about human history [2, 3, 4, 5, 6, 7, 8].

Genetic variation is useful in understanding all the above as the genetic information of an individual carries with it the history of her ancestrals. In particular, it has been shown several times that genetic variation encodes with high accuracy an individual's geographic origin even in closely related populations such as the European populations [9].

The information obtained from the DNA, the genotypes of a sample of individuals from a population, is a multidimensional data which can be thought of as a matrix of individuals and SNPs. Principal component analysis (PCA) is typically perforemed in order to relate the individual's genotype to its geographic location. PCA projects information from hundreds of thousands of genetic variants onto two dimensions which maximize the observed variance. In [9] it is shown that the projections of individuals on these principal components mirror the locations of their ancestral populations with a remarkably high resolution. The underlying cause of this observation is that genetic variation encodes spatial structure [5]. That is, individuals originating close to each other tend to have more similar genetic variation than individuals who originate from different regions.

Prinicipal Component Analysis is a natural choice for the analysis of genotype data, as it is a classical statistical tool for dimension reduction, which has been well studied in a number of areas, including statistics, machine learning and pattern recognition. Mathematically, PCA looks for the low-rank approximation to the covariance matrix among all the samples. While PCA remarkably captures the locations of individuals, the principal components themselves only indirectly capture the spatial structure of genetic variation, and it lacks direct model-based interpretation. In this paper we propose an alternate, more natural approach for modeling spatial structure of genetic variation where we explicitly model how the allele frequency of each SNP changes as a function of the location of the individual in space (i.e., the allele frequency is a

function of the (x, y) coordinates of an individual in the map. We derive an Expectation-Maximization algorithm that simultaneously estimates the position-specific allele frequencies of all SNPs, and the (x, y) map-coordinates of each genotype in the sample.

We demonstrate the utility of our model in a few scenarios. First, by utilizing the fact that our model provides an explicit representation of the allele frequency as a function of the map coordinates, we demonstrate that the model can predict the geographical origins of an individual even in the case that the individual is of mixed ancestry (e.g., Spanish mother and Swedish father); in contrast to PCA, which will map such an individual to central Europe, our approach maps the two parents directly. Second, we detect regions under selection by searching for SNPs in which the gradient of the allele frequency function differs significantly from the average gradient across all SNPs. Our prediction for regions under selection is consistent with previous methods for selection detection, such as iHS [10] and Fst [11], and regions under selection such as the LCT and HLA regions stand out clearly using our method. In contrast to previous methods, however, our method does not require a homogeneous population (such as in the case of iHS), or a dichotomous parition of the population into subpopulations (e.g., in Fst analysis). This allows us to detect additional signals of selection across the genome that were not detected previously. Finally, we extend our approach to model spatial structure over the unit sphere to predict the spatial structure of worldwide populations.

[1] A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich, "Principal components analysis corrects for stratification in genome-wide association studies," Nature Genetics, vol. 38, no. 8, pp. 904–909, July 2006.
[2] P. Menozzi, A. Piazza, and L. Cavalli-Sforza, "Synthetic maps of human gene frequencies in europeans," Science, vol. 201, no. 4358, pp. 786–792, September 1978.
[3] L. L. Cavalli-Sforza, P. Menozzi, and A. Piazza, The history and geography of human genetics. Princeton, New Jersey, USA: Princeton University Press, 1994.
[4] L. L. Cavalli-Sforza, "The human genome diversity project: past, present and future," Nature Review Genetics, vol. 6, no. 4, pp. 333–340, April 2005.
[5] N. A. Rosenberg, J. K. Pritchard, J. L. Weber, H. M. Cann, K. K. Kidd, L. A. Zhivotovsky, and M. W. Feldman, "Genetic structure of human populations," Science, vol. 298, no. 5602, pp. 2381–2385, December 2002.
[6] M. Bamshad, S. Wooding, B. A. Salisbury, and J. C. Stephens, "Deconstructing the relation- ship between genetics and race." Nature Review Genetics, vol. 5, no. 8, pp. 598–609, August 2004.
[7] J. Z. Li, D. M. Absher, H. Tang, A. M. Southwick, A. M. Casto, S. Ramachandran, H. M. Cann, G. S. Barsh, M. Feldman, L. L. Cavalli-Sforza, and R. M. Myers, "Worldwide human relationships inferred from genome-wide patterns of variation," Science, vol. 319, no. 5866, pp. 1100–1104, February 2008.
[8] M. Jakobsson, S. W. Scholz, P. Scheet, R. J. Gibbs, J. M. Vanliere, H.-C. Fung, Z. A. Szpiech, J. H. Degnan, K. Wang, R. Guerreiro, J. M. Bras, J. C. Schymick, D. G. Hernandez, B. J. Traynor, J. Simon-Sanchez, M. Matarin, A. Britton, J. van de Leemput, I. Rafferty, M. Bucan, H. M. Cann, J. A. Hardy, N. A. Rosenberg, and A. B. Singleton, "Genotype, haplotype and copy-number variation in worldwide human populations," Nature, vol. 451, no. 7181, pp. 998– 1003, February 2008.
[9] J. Novembre, T. Johnson, K. Bryc, Z. Kutalik, A. Boyko, A. Auton, A. Indap, K. King, S. Bergmann, M. Nelson, M. Stephens, and C. Bustamante, "Genes mirros geography within europe," Nature, vol. 456, pp. 98–101, 2008.
[10] B. F. Voight, S. Kudaravalli, X. Wen, and J. K. Pritchard, "A map of recent positive selection in the human genome," PLoS biology, vol. 4, pp. e72+, 2006.
[11] K. E. Holsinger and B. S. Weir, "Genetics in geographically structured populations: defining, estimating and interpreting fst," Nature Reviews Genetics, vol. 10, pp. 639–650, 2009.

New Algorithm for Constructing Supernetworks from Partial Trees

Changiz Eslahchi, Reza Hassanzadeh
Faculty of Mathematics, Shahid Beheshti University, G.C., Tehran, Iran, ch-eslahchi@sbu.ac.ir

The evolution of a group of taxa or genes is generally perceived as a single tree. However, Analyses based on data from different genes return different trees. Moreover, different analysis on a data also result different trees. So, phylogeneticists often confront with several different trees and want to summarize the collections of partial phylogenetic trees in the form of supertrees or supernetworks. In this paper, we introduce a new algorithm based on simulated annealing for constructing phylogenetic supernetworks from partial trees. Before describing the algorithm, first we summarize some necessary concepts.

Suppose that X is a finite set of taxa. A *split* $A|B$ of X is a bipartition of X, i.e., a partition of X into two non-empty sets or parts A and B with $A \cup B = X$ and $A \cap B = \phi$. A *split* $A|B$ is called *partial split* if $A \cup B \subset X$. A *phylogenetic tree* (on X) is a tree with leaves labelled by X. Each edge of a phylogenetic tree naturally gives rise to a split. A collection of splits of X is *circular* if it has the following property: There exists an ordering $x_1, \dots, x_n$ of X such that every split is of the form $\{x_i, x_{i+1}, \dots, x_j\} | X - \{x_i, x_{i+1}, \dots, x_j\}$ for some i and j, $1 \leq i \leq j \leq n$. We say that a split is realized by a circular ordering if it has above property. A. Dress and D. H. Huson proved that circular collections of splits always have a planar splits graph representation [1].

Now, suppose a collection of partial trees is given. Our algorithm, SNSA (Super Network Simulated Annealing), extracts all partial splits from partial trees. Then it tries to find a circular ordering based on simulated annealing such that the number of realized partial splits is maximal. After producing circular ordering, the collection of splits obtained from the circular ordering are weighted using non-negative least squares. Then, we use SplitsTree4 program [2] to draw supernetworks from these weighted splits.

To illustrate the applicability of SNSA, we performed it on 2 biological datasets, which are available in splitstree4 program, and compared the outputs with Z-Closure [3]. Figures 1 and 2 show the supernetworks obtained by both algorithms. For small dataset, 2galles, the outputs of both algorithms are the same (Figure 1). But for large dataset, fungi, the results are different (Figure 2). According to what we see in figure 2, it seems that both algorithms give the same

classification of taxa and exhibit the same major splits. But the supernetwork obtained by SNSA is simple and less complicated than supernetwork obtained by Z-Closure. One of the main advantages of SNSA is that it produces a planar network. So, analysis of supernetworks obtained by SNSA is more comfortable than supernetworks obtained by Z-closure.

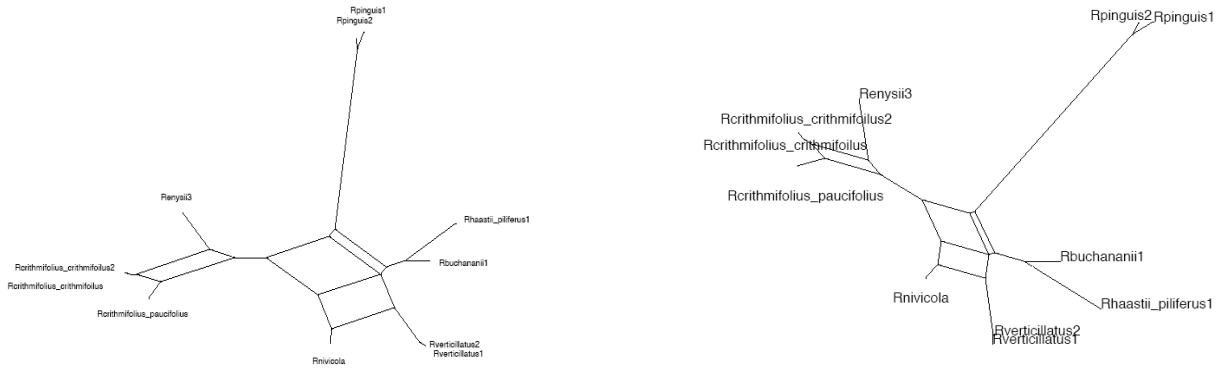


Figure1. The SNSA network (Left) and the Z-Closure network (Right) for the 2galles data set.

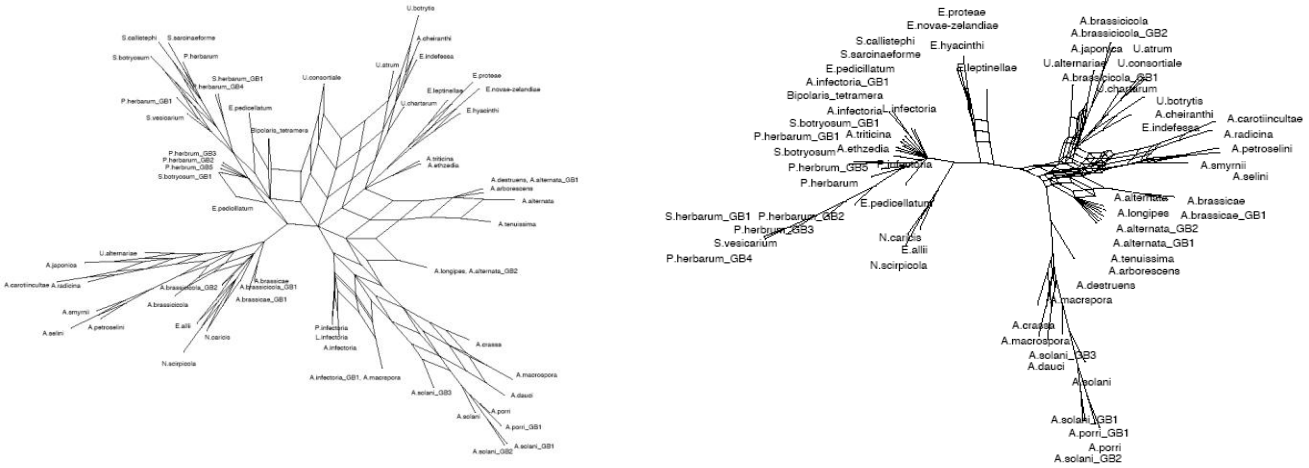


Figure2. The SNSA network (Left) and the Z-Closure network (Right) for the fungi data set.

1. A. Dress, D.H. Huson (2004) Constructing splits graphs, *IEEE/ACM Transactions in Computational Biology and Bioinformatics*, **1:**109–115.
2. D.H. Huson, D. Bryant (2005) Application of phylogenetic networks in evolutionary studies, *Adv. Math*, **92:**47–105.
3. D.H. Huson et al. (2004) Phylogenetic super-networks from partial trees, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **1:**151–158.

Molecular phylogeny analysis in cotton

Farah Farahani[1], Masoud Sheidai[2]

[1]Islamic Azad university, Iran; [2]Shahid Beheshti university, Iran, farahfarahani2000@yahoo.com

Cotton is an important economic and fiber crop, grown in 70 countries in the world. Over 180 million people are associated with the fiber industry that produces 20 to 30 billion dollars worth of raw cotton. Molecular phylogeny analysis of some tetraploid (Gossypium hirsutum) cotton cultivars and their hybrids were performed by using RAPD molecular markers. DNA extraction was done by using the CTAB method (Murry and Tompson 1980) with modification described by De La Rosa et al. (2002). Out of 30 primers used 22 primers produced reproducible bands. In total 387 bands (loci) were obtained out of which 178 bands were polymorph and only 209 bands were common in cotton cultivars studied. These common bands are sympleisomorphic loci, the conserved sequences in cotton genome. Some of the cultivars showed presence of specific bands which are synapomorphic loci obtained during genetic differentiation. For example band no. 13 (1550 bp), of the primer OPM011 was specific for the hybrid cultivar Nazilli X Tabladilla, band No. 1 (700 bp) of the primer OPB07 was specific for the hybrid cultivar Siokra X Tabladilla, band No. 2 of the primer OPH07 (320 bp) was specific for the cultivar Nazilli while band No. 5 of the same primer occurred only in the cultivar Tabladilla. Some bands were present in all genotypes but absent only in one. This happens due to loss of genetic material n the genome. For example bands N0. 5-7 (750, 800 bp respectively) of the primer OPM11 were absent only in the hybrid cultivar Siokra X Tabladilla. Band No. 6 (1100 bp) of the primer OPH19 was absent only in the hybrid cultivar Nazilli X Tabladilla. Some bands were absent in the parental genotypes but present only in their hybrids, whereas some bands occurred in the parental genotypes but absent in their hybrid, all these indicate genetic recombination during hybridization. For example bands No. 9 and 11 of the primer OPA- 11 (2200 & 2900 bp respectively) occurred in the hybrid cultivar Sahel X Siokra but were absent in its parental genotypes. Bands No. 20 and 25 of the primer OPM-11 (2600 & 1000 bp respectively) were present in the parental genotypes of Siokra and Nazilli but were absent in their hybrid. NJ and Bayesian clustering based on Nei genetic distance (Nei 1972), grouped two parental cultivars of Nazilli and Siokra together with 100% bootstrap value, these genotypes show genetic difference with the others and stand separate from them. This is also true for Sahel genotype as it stands far from all other genotypes. However, genetic affinity was observed between hybrid cultivars of Sahel X Siokra and Nazilli X Tabladilla and also between Sahel X Nazilli and Sahel X Tabladilla. Genetic relationship between other parental and hybrid genotypes will be discussed in detail in the paper. Such molecular phylogeny analysis helps us to plan a better breeding and hybridization program in cotton.

1- R. De la Rosa, C. James, Tobutt K.R (2002) Isolation and characterization of polymorphic microsatellite in olive (Olea europaea L.) and their transferability to other genera in the Oleaceae. Primer Note. Mol Ecol Notes, 2: 265-7.
2- MG. Murry, W.F. Tompson (1980) Rapid isolation of high molecular weight plant DNA. Nucleic Acids Res, 8: 4321-4326.
3- M. Nei (1972) Genetic distance between populations. Am Nat, 106: 283-92.

APSampler: open-source software for identifying multigene effects in genetic data

Alexander Favorov[1,2,3], Dmitrijs Lvovs[2], Marina Sudomoina[4], Olga Favorova[4], Giovanni Parmigiani[5], Michael F. Ochs[6,7]

[1]*Johns Hopkins University, Baltimore, MD, USA*
[2]*Research Institute for Genetics and Selection of Industrial Microorganisms, Moscow, Russia*
[3]*Vavilov Institute of General Genetics, Moscow, Russia*
[4]*Department of Molecular Biology, Russian State Medical University, Moscow, Russia*
[5]*Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, MA, USA*
[6]*Department of Oncology, School of Medicine, Johns Hopkins University, Baltimore, MD, USA*
[7]*Department of Health Science Informatics, School of Medicine, Johns Hopkins Univ, Baltimore, MD, USA*
favorov@sensi.org

Previously we developed the APSampler (1) algorithm, which identifies allelic patterns that are associated with a disease in an individual genotype-phenotype dataset, in order to aid in the deciphering of the genetic picture in multiallelic diseases. While our APSampler algorithm has proven to be useful and reliable in genetic disease studies over the last five years (2,3), it did not include tools for proper statistical validation and was difficult for users to apply due to its data format and lack of convenient code resources.

Here we present an open-source release of APSampler that uses human-readable and generatable genetic data and includes a statistical framework to validate identified genetic interactions. The new open-source code, User's Manual, and statistical framework will provide geneticists and biologists working with genotypic data a proven tool for discovery of genetic interactions.

APSampler is a reliable and sensitive heuristic tool that identifies multiallelic signatures associated with a diseases and that validates the findings. The APSampler software package is available as C source code for the main algorithm together with a set of Perl scripts that implement the validation framework. The source code is prresented at http://www.cancerbiostats.onc.jhmi.edu/APSampler.cfm; the Open Source APSampler project is available under the MIT Artistic license at http://code.google.com/p/apsampler/.

1. A.V.Favorov et al (2005) A Markov chain Monte Carlo technique for identification of combinations of allelic variants underlying complex diseases in humans., Genetics, 171(4):2113-21.
2. C.O'Doherty et al (2009) Genetic polymorphisms, their allele combinations and IFN-beta treatment response in Irish multiple sclerosis patients, Pharmacogenomics, 10(7):1177–86.
3. M.A.Sudomoina et al (2010) [Complex analsis of association of inflammation genes with myocardial infarction] (rus) Mol Biol (Mosk) 44(3):463-71.

An Internet Service On A Folding Energy Estimate

Sergey Feranchuk[1], Alexander Tuzikov[1], Dmitry Mukha[2], Ulyana Potapova[3],

[1]*United Institute of Informatics Problems NAS Belarus, Belarus, feranchuk@gmail.com*
[2]*Institute of Bioorganic Chemistry NAS Belarus, Belarus, dvmukha@gmail.com*
[3]*Limnological Institute SB RAS, Irkutsk, shuana1983@yandex.ru*

The internet service presented on a web-site http://bri-shur.com is designed as a tool of a pipeline on protein structure prediction. To build a protein structure with homology modeling, typically one needs to find a correct template with a known structure, build a correct pairwise alignment between a query sequence and a template and build the structure using the alignment and template coordinates. All these steps are supported by presented web services. The homology screening in PDB almost always allows to find one or several templates. However, when there are no close homologs for a query sequence, there can be a choice between several templates and even between several fold classes. For the last case, one needs to use additional reasons to choose a correct fold type for the query sequence.

For that purpose the service on folding energy estimate for a given structure is implemented. Hydrophobic energy and the energy of electrostatic interactions, as it is commonly accepted [1], make a major contribution to a free energy of a protein structure. These terms are are compensated by the loss of a conformational entropy of a folded structure. The correct estimation for a conformational entropy seems to be unavailable in a reasonable time, so the service gives only the values of electrostatic energy and hydrophobic energy; however to account the last term, some average level is subtracted from both energies to make difference between a correct and incorrect folds more clear.

The results of predictions accomplished by means of the presented web-service are summarized at the table below, where the energy values for a correct fold and for a decoy are compared. The decoys were taken as the closest match in a homology screening, next to matches to a correct fold. Units of the energy are somewhat arbitrary, because the estimate is too rough to declare it as a correct energy. However the normalizing factors are chosen so that an estimate for a correct structure has a similar value as an experimental folding energy.

Protein	1	2	3	4	5	6	7
BPTI	**-1.35**	-0.86	-0.36	**-1.22**	-0.47	-0.23	**-0.70**
Barnase	**-1.59**	-1.23	-0.42	**-1.65**	-0.38	-0.22	**-0.60**
Myoglobin	**-1.66**	-1.09	-1.38	**-2.47**	0.33	-0.31	**0.02**
Lysozyme	**-1.73**	-1.28	-0.32	**-1.6**	-0.33	-0.12	**-0.45**
Cytochrome c	**-1.50**	-1.23	-0.53	**-1.76**	-0.90	-0.33	**-1.23**
Ubiquitin	**-1.66**	-1.00	-0.32	**-1.32**	-0.90	-0.36	**-1.26**

1 - experimental enthalpy of folding, according to [2], in kcal/mol
2,3,4 - hydrophobic energy, electrostatic energy and total energy estimate for a correct model
5,6,7 - hydrophobic energy, electrostatic energy and total energy estimate for a decoy

References:

1. R.L.Baldwin (2007) Energetics of protein folding, *J Mol Biol*, **371:**283-301.
2. P.L.Privalov et al. (2007) Microcalorymetry of biological macromolecules, *Biophys Chem*, **126:**16–24.

Potential function of proteins encoded by chimeric transcripts

Milana Frenkel-Morgenstern, Iakes Ezkurdia, David Pisano,
Angela Del Pozo, Michael Tress and Alfonso Valencia

Spanish National Cancer Research Centre (CNIO), Madrid, 28029, Spain,
mmorgenstern@cnio.es, avalencia@cnio.es

Chimeric RNAs are distinct from conventional alternatively spliced isoforms [1-11], they may result from the trans-splicing of pre-mRNAs, from the "polymerase jumping" between different regions of RNAs to form a single pre-mRNA, or alternatively, may be the product of gene fusion following translocations or rearrangements [3, 8, 12]. Chimeric transcripts can contribute to the complex completion of distinct cellular processes, permitting a combinatorial increase in the gene products available [8, 13]. So far, only a limited number of chimeric transcripts and/or their associated protein products have been characterized, most of which result from chromosomal translocations and many of which are associated with cancer [12, 14-18]. Therefore, it is important to extend these observations in order to catalog the chimeric transcripts that exist and to study the potential functions of their corresponding chimeric proteins.

Using the paired-end transcriptome sequencing approach, we confirmed the existence of 53% of the chimeric transcripts that were initially detected by the screening of EST libraries in human, mouse, and fruit fly [19]. In addition, we report that a small but significant number of

these transcripts are translated into proteins by detecting the presence of unique peptides that matched the gene-gene junctions of the chimeras. This identification required intensive searches of mass-spectrometry databases [20] and in some cases, there was evidence that the corresponding chimeric proteins might be distributed in a tissue-specific manner.

A number of the protein products potentially generated from such chimeric transcripts contain complete functional domains, in many cases associated with transmembrane domains and signal peptides. The persistence of these structures might indicate that new functions can be acquired by combining autonomous protein domains. Indeed, more than 12% of the chimeras reported represent proteins with complete DNA binding domains but that lack the activation domains of transcription factors, or enzymes that lack the sites required for the formation of functional dimers, indicating that these proteins might have dominant negative effects. Intriguing examples of the possible dominant-negative effect of chimera of E2-alpha transcription factor, TCF3, and the C-terminal binding protein-1, CTBP1, in mouse, and MLL/GMPS fusion protein in human are discussed.

In summary, the evolutionary constraints imposed by the linear arrangement of exons in a given gene can be overcome through distinct chromosomal rearrangements. By analyzing the consequences of translating these chimeric transcripts into proteins, it appears that these novel chimeric proteins are likely to have substantially different functions to the original native proteins. Indeed, they may be found in distinct cellular compartments or in specific tissues, perhaps associated with specific diseases and cancers. Furthermore, it seems feasible that these chimeras could have acquired specific functions and that they could have dominant negative effects due to the absence of certain functional domains, actively competing with the functional wild type proteins. Significantly, a number of examples already point in this direction.

1. Horiuchi, T. and Aigaki, T. (2006) *Biol Cell* 98, 135-140
2. Robertson, H.M., *et al.* (2007) *Genetics* 176, 1351-1353
3. Herai, R.H. and Yamagishi, M.E. (2010). *Brief Bioinform* 11, 198-209
4. Douris, V., *et al.* (2010). *Mol Biol Evol* 27, 684-693
5. Pettitt, J., *et al.* (2010). *Biochem Soc Trans* 38, 1125-1130
6. Allen, M.A., *et al.* (2010). *Genome Res*
7. McManus, C.J., *et al.* (2010). *Proc Natl Acad Sci U S A* 107, 12975-12979
8. Gingeras, T.R. (2009). *Nature* 461, 206-211
9. Li, H., *et al.* (2008) . *Science* 321, 1357-1361
10. McManus, C.J., *et al.* (2010). *Genome Res* 20, 816-825
11. Pirrotta, V. (2002) Trans-splicing in Drosophila. *Bioessays* 24, 988-991
12. Maher, C.A., *et al.* (2009). *Proc Natl Acad Sci U S A* 106, 12353-12358
13. Birney, E., *et al.* (2007). *Nature* 447, 799-816
14. Silberg, J.J., *et al.* (2010). *Methods Mol Biol* 673, 175-188
15. Candel, A.M., *et al.* (2009). *Protein Eng Des Sel* 22, 597-606
16. Miura, T.A., *et al.* (2004). *J Virol* 78, 4646-4654
17. Eguchi, M., *et al.* (2006). *Genes Chromosomes Cancer* 45, 754-760
18. Mitani, K. (2004). *Oncogene* 23, 4263-4269
19. Li, X., *et al.* (2009). *J Mol Evol* 68, 56-65
20. Tress, M. *et, al.* (2010) Naturally occuring trans-spliced protein isoforms identified in *Drosophila*. Submitted.

The role of recombination in the multiplication of Alu-associated microsatellites

Marina Fridman[1], Nina Oparina[2], Ivan Kulakovskiy[1], Vsevolod Makeev[1]

[1]*Vavilov Institute of General Genetics, RAS, Moscow, Russian Federation, marina-free@mail.ru*
[2]*Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Russian Federation*

Previously we have demonstrated that frequent repeats in mammal genomes are generated by active retroposons (3). In fact despite the importance of microsatellites for genome function, (de)stabilization etc. the repertoire of frequent types of microsatellites in mammalian genomes is limited: most of monomers are AT-rich motifs like (A)nB. What is the source of these repeats? We suppose that this is poly-A tail of Alu's and L1's. We have demonstrated that the vast majority of these repeats is associated with 3'-ends of retroposons (Alu and L1 in the human genome, B1 in mouse genome etc.). Several cases of other associations were found in cow genome and mouse genome (LTR association and undescribed Y-chromosome specific repetitive element). The common feature of these cases: both LTR/Y-repeat and associated microsatellites are located within genomic duplications. Thus we have proposed that "seeds" of microsatellites are located in poly-A tails of retroposons. Less A-rich microsatellites could be generated from other "seeds". But is their generation associated with retroposon integration or occurs later? We have investigated processed pseudogens and the vicinity of their poly-A. They use for transposition the same mechanism that L1 and Alu do. But there are a few microsatellites on theirs flanks. Retroposons are the loci of active recombination and conversion though meiotic recombination near Alu's is rare (2). Thus, it is necessary to look for other processes responsible for recombination. Aleshin and Zhi (1) found sequence homogenization of neighboring Alu elements. They suppose that such homogenization is the signature of nonallelic gene conversion. We found the signature of nonallelic gene conversion for microsatellites near neighboring Alu's. Thus, gene conversion may contribute in the generation of microsatellites or theirs «seeds».

It is noteworthy that most pseudogenes that have tandem repeats on theirs flanks have also many copies in genome. So, they may be the subjects of conversion.

1. A.Aleshin and D.Zhi (2010) Recombination-associated sequence homogenization of neighboring Alu elements: signature of nonallelic gene conversion, Mol Biol Evol 27(10): 2300-11.
2. S.Dadashev et al (2005) In silico identification and characterization of meiotic DNA: AluJb possibly participates in the attachment of chromatin loops to synaptonemal complex,
Genetika 41(12):1707-13.
3. M.Fridman et al (2010) Frequent repeats in mammal genomes and active retroposones. Proceedings of the 7th Int. Conf. on Bioinformatics of Genome Regulation and Structure, Novosibirsk.

Exploring the fold space of membrane proteins

Dmitrij Frishman

Technische Universität München, Maximus-von-Imhof-Forum 3, 85354, Freising, Germany,
d.frishman@wzw.tum.de

Recent progress in structure determination techniques has led to a significant growth in the number of known membrane protein structures, and the first structural genomics projects focusing on membrane proteins have been initiated, warranting an investigation of appropriate bioinformatics strategies for optimal structural target selection for these molecules. What determines a membrane protein fold? How many membrane structures need to be solved to provide sufficient structural coverage of the membrane protein sequence space? In my talk I will describe the CAMPS database (Computational Analysis of the Membrane Protein Space) that automatically classifies a-helical membrane proteins into fold classes. I will also present our latest results on predicting interacting helices in membrane proteins and deriving helix connectivity diagrams that represent a convenient level of abstraction for comparing predicted membrane protein structures. Finally, I will review the difficulties in classifying experimentally determined three-dimensional structures of membrane proteins, focusing on difference between the SCOP and CATH databases.

Frishman, D. (Ed) Structural bioinformatics of membrane proteins. Springer, Wien 2009. ISBN 978-3709100448.
Fuchs A, Frishman D. (2010) Structural comparison and classification of alpha-helical transmembrane domains based on helix interaction patterns. *Proteins*, **78**, 2587-2599.
Neumann S., Fuchs A., Mulkidjanian A, and Frishman D. (2010) Current status of membrane protein structure classification. *Proteins*, **78**, 1760-1773.
Fuchs, A., Kirschner, A., Frishman D. (2009) Prediction of helix-helix contacts and interacting helices in polytopic membrane proteins using neural networks. *Proteins*, **74**, 857-71
Fuchs, A., Martin-Galiano, A.J., Kalman, M., Fleishman, S., Ben-Tal, S., and Frishman, D. (2007) Co-Evolving Residues in Membrane Proteins. *Bioinformatics*, **23**, 3312-3319.
Martin-Galiano, A.J., Frishman, D. (2006) Defining the Fold Space of Membrane Proteins: the CAMPS Database. *Proteins*, **64**, 906-22.

Influence of organization of native structure on its folding: Modeling of protein folding

Oxana Galzitskaya[1], Natalya Bogatyreva[1], Anna Glyakina[2]

[1]*Institute of Protein Research RAS, Russian Federation, ogalzit@vega.protres.ru*
[2]*Institute of Mathematical Problems of Biology RAS, Russian Federation, glyakina@rambler.ru*

The problem of protein self-organization is one of the most important problems of molecular biology nowadays. Despite the recent success in the understanding of general principles of protein folding, details of this process are yet to be elucidated. Moreover, the prediction of protein folding rates has its own practical value due to the fact that aggregation directly depends on the rate of protein folding. The time of folding and transition state ensembles for 67 proteins with known experimental data at the point of thermodynamic equilibrium between unfolded and native state have been calculated using a Monte Carlo method and Dynamic programming one where each residue is considered to be either folded as in the native state or completely disordered. The times of folding for 67 proteins which reach the native state within a limit of 108 Monte Carlo steps are in a good correlation with experimentally measured folding time at mid-transition point (the correlation coefficient is -0.82). A lower correlation was obtained if to use Dynamic programming approach (the correlation coefficient is -0.72). The capillarity model allows us to predict the folding rate at the same level of correlation as by Monte-Carlo simulations. The calculated model entropy capacity (conformational entropy per residue divided by the average contact energy per residue) for 67 proteins correlates by about 78% with the experimentally measured folding rate at the mid-transition point. Theoretical consideration of a capillarity model for the process of protein folding demonstrates that the difference in the folding rate for proteins sharing more ball-like and less ball-like folds is the result of differences in the conformational entropy due to a larger surface of the boundary between folded and unfolded phases in the transition state for proteins with more ball-like fold.

An important question that is touched upon here is whether the modeling of protein folding can catch the difference between the folding of proteins with similar structures but with different folding mechanisms. The modeling of folding of six alpha-helical proteins, which are similar in size, was done by using the Monte Carlo and dynamic programming methods. A frequently observed order of folding of alpha-helices for each protein was determined using the Monte Carlo method. A correlation between the experimental folding rate and the number of Monte Carlo steps was also demonstrated for right-handed proteins. Amino acid residues which

are important for the folding were determined using the dynamic programming method. These defined regions correlate with the order of folding of secondary-structure elements in these proteins both in experiments and in modeling.

Statistical analysis of protein folding rates has been done for 84 proteins with available experimental data. A surprising result is that the proteins with multi-state kinetics from the size range of 50-100 amino acid residues fold as fast as proteins with two-state kinetics from the same size range. At the same time, the proteins with two-state kinetics from the size range 101-151 fold faster than those from the size range 50-100. Statistical analysis of folding rates of 73 proteins with known experimental data revealed that bacterial proteins with simple kinetics (23 proteins) exhibit a higher folding rate compared to eukaryotic proteins with simple folding kinetics (27 proteins).

This work was supported by the programs "Molecular and cellular biology" and “Fundamental sciences – medicine”, by the Russian Foundation for Basic Research.

"Golden Triangle" for Protein Folding Rates

Sergiy Garbuzynskiy, Dmitry Ivankov, Natalya Bogatyreva, Alexei Finkelstein

[1]*Institute of Protein Research, Russian Academy of Sciences, Russian Federation,*
sergey@phys.protres.ru, afinkel@vega.protres.ru

A decade ago, it has been theoretically shown[1,2] that: (1) the folding rate of a single-domain globular protein, in a point of equilibrium of its native and denatured states, must fall between $10^8 s^{-1} \times \exp(-0.5L^{2/3})$ and $10^8 s^{-1} \times \exp(-1.5L^{2/3})$, where $10^8 s^{-1}$ is the experimentally measured rate of conformational rearrangement of one amino acid residue, and L is the number of residues in the protein chain; (2) if ΔG is the free energy difference between the native and denatured states of a protein, then its folding rate increases by about $\exp(-\Delta G/2RT)$ times, where T is temperature and R is the gas constant. Besides, for obvious biological reasons, protein folding rates cannot be smaller than $\sim 10^{-2} s^{-1} \div 10^{-3} s^{-1}$. Taken together, these three limits (two physical, one biological) outline a theoretically allowed "golden triangle" for folding rates of single-domain globular proteins of any size and stability. In this communication, we show that the experimentally determined protein folding rates indeed fall within this "golden triangle". Besides, we predict that (1) the upper limit of the size of a protein domain folding under complete thermodynamic control is about 80 amino acid residues, while the upper limit of the size of a domain capable of folding is about 600 amino acid residues. The collected statistics on protein spatial structures confirms both these predictions.

This work was supported by the grants from Howard Hughes Medical Institute (#55005607), RFBR (#10-04-00162a), MCB (#01200957492) and "Leading Scientific Schools" (#NSh-2791.2008.4) programs, FASI (#02.740.11.0295), by grant of the Dynasty Foundation and the Russian Young Scientists' grants (#MK-4894.2009.4, #MK-5540.2011.4).

1. A.V.Finkelstein, A.Y.Badretdinov (1997) Rate of protein folding near the point of thermodynamic equilibrium between the coil and the most stable chain fold, *Fold. & Des.*, **2:**115–121.
2. A.V.Finkelstein (2003) Proteins: thermodynamic, structural and kinetic aspects, In: *Slow relaxations and nonequilibrium dynamics in condensed matter*, J.-L.Barrat et al. (Eds.), 649–704 (Springer-Verlag: Berlin, New York).

Horizontal gene transfer and genome evolution in Methanosarcina

Sofya Garushyants, Marat Kazanov
IITP RAS, Russian Federation, garushyants@gmail.com

Methanosarcina spp. are characterized by the largest in the Archaea domain genomes. Genome size of this genus is considered to be the result of high level of horizontal gene transfer (HGT) from Bacteria. About 30 % of genes were shown to be transferred from Bacteria, these were involved in central metabolism and solute transport, such as sugar synthesis, sulfur metabolism, phosphate metabolism, DNA repair, transport of small molecules and others [1]. Less is known about how all these genes are dispersed along the Methanosarcina genomes, and how they are regulated. Are these genes situated in clusters or are they inserted in typical Archaea operons? Are these genes regulated by archaeal-like type transcription regulators or by bacterial-like?

To answer all these questions comparative genomics analysis was carried out. All orthologous groups, that included genes from the three known Methanosarcina species were identified and 460 genes from 147 orthologous groups were shown to be bacterial-like. For these selected genes genome context was analyzed, and it was shown, that HGT genes tended to be included in operons not only with other HGT genes, but with archaeal-like genes as well. Phylogenetic analysis revealed that most laterally transferred genes in Methanosarcina belong to Clostridia, or Bacilli, or various Proteobacteria; a few Cyanobacteria like genes were also found. Our data showed that the amount of transcriptional regulators among HGT genes was lower than the average amount of transcription regulators per Methanosarcina genome, besides that no conservative bacterial-like regulation signals were found in upstream areas of the operons studied, so it was presumed that bacterial-like genes are regulated mostly by archaeal transcription factors. Also it was shown, that the HGT effect on Methanosarcina genome size was overestimated in previous researches. Additionally all ortologous groups for all known Methanosarcinaceae were analyzed and it was shown, that HGT events occurred between Bacteria and common ancestor of all modern Methanosarcinaceae. This result corresponds to the data on archaeon Methanococcoides burtonii [2].

This work in made in cooperation with Mikhail Gelfand (IITP RAS).

1. Deppenmeier U, Johann A, Hartsch T, Merkl R, Schmitz RA, Martinez-Arias R, Henne A, Wiezer A, Bäumer S, Jacobi C, Brüggemann H, Lienard T, Christmann A, Bömeke M, Steckel S, Bhattacharyya A, Lykidis A, Overbeek R, Klenk HP, Gunsalus RP, Fritz HJ, Gottschalk G.(2002) The genome of Methanosarcina mazei: evidence for lateral gene transfer between bacteria and archaea. J Mol Microbiol Biotechnol. 453-461.
2. Allen MA, Lauro FM, Williams TJ, Burg D, Siddiqui KS, De Francisci D, Chong KW, Pilak O, Chew HH, De Maere MZ, Ting L, Katrib M, Ng C, Sowers KR, Galperin MY, Anderson IJ, Ivanova N, Dalin E, Martinez M, Lapidus A, Hauser L, Land M,Thomas T, Cavicchioli R. (2009) The genome sequence of the psychrophilic archaeon, Methanococcoides burtonii: the role of genome evolution in cold adaptation. ISME J. 1012-1035.

Study of DNA Binding Proteins in E. coli and their role in organization of nucleoid structure

Payel Ghosh[1], Debashree Basu[2], Shubhada R. Hegde[1], Shekhar C. Mande[1]

[1]*Centre for DNA Fingerprinting & Diagnostics, India, payel@cdfd.org.in*
[2]*University of Texas Medical Branch, United States, debasu@utmb.edu*

Unlike eukaryotes, some of the basic elements of DNA compaction such as histones are absent in bacteria, and therefore exact molecular mechanisms of bacterial chromosome packaging are still unclear. Among the factors facilitating DNA condensation, binding of nucleoid associated proteins (NAPs) may play a major role in prokaryotic genome organization. The aim of the study is to identify different sites on the *Escherichia coli* genome where the NAPs might bind and how their association helps to form a compact nucleoid structure. While most of the previous efforts in this direction were confined to investigating the binding patterns of a specific group of NAPs, ours was a non-specific approach – aimed at determining all possible DNA-protein interactions. The complete genome of the *E. coli* K12 isolate MG1655 was reviewed for DNA-protein binding sites, with the help of 'ChIP-chip' experiments. The locations of protein binding sites thus detected were compared with previously known data for specific NAPs, and also with putative binding sites identified *in-silico* by scanning the DNA for specific 'consensus' sequence/motifs.

The experiments have been repeated after subjecting the *E. coli* cells to hyperosmotic stress (imposed by high concentrations of sucrose). It is well-known that bacterial cells like *E. coli* can control their own hydration by accumulating solutes when they are exposed to media having high osmolarity, and by releasing solutes in response to osmotic down-shocks. Comparison of results obtained for different *E. coli* cultures indicate if-and-how the binding patterns of some NAPs change in course of cellular osmoregulation. The reported trends in binding of NAPs are expected to better reveal the structural organisation and dynamicity of the *E. coli* nucleoid.

Understanding aging through genome analysis

Vadim Gladyshev

Brigham and Women's Hospital, Harvard Medical School, United States
vgladyshev@rics.bwh.harvard.edu

Evolution of bacterial pan-genomes

Evgeny Gordienko[1], Marat Kazanov[2], Mikhail Gelfand[2]

[1]*N.I. Vavilov Institute of General Genetics RAS, Russian Federation, egordienko@vigg.ras*
[2]*Institute for Information Transmission Problems of the Russian Academy of Sciences, Russian Federation*

Analysis of completely sequenced bacterial genomes shows high variability of their size and content, even for closely related strains. For instance, only about 39% of proteins are common for three strains of *E.coli* [1].

As a result, specie definition based on a genome of a single strain is insufficient. For the description of the bacterial "specie genome", based on the sequences of different strains, the "pan-genome" approach was suggested [2]. The pan-genome is a total complement of genes from all strains of the same specie (or genus). The pan-genome consists of three parts: the universal genome with genes common for all strains; the unique genome with strain-specific genes; and the periphery (genes present in a subset of strains).

We performed pairwise comparison of pan-genomes of the genera *Escherichia*, *Shigella* and *Salmonella* from the family *Enterobacteriaceae*. At the initial step, the orthology table was constructed using a BLASP-based algorithm. We filtered out all proteins linked to transposons, IS elements or phages.

Pan-genomes were compared, and their components were quantified and analyzed by GOstat. We normalized for the different number of genomes in the genera, and performed bootstrapping to quantify robustness and saturation of the observations.

Unlike previous studies, we performed pairwise comparison of pan-genomes. For each gene, the fraction of genomes from two genera (for instance, *Escherichia* and *Salmonella*) containing this gene was determined. A plane chart was designed to visualize such comparisons. The common universal genomes for both genera contain mainly housekeeping genes. The unique genomes contain many flagellum-linked genes.

Unexpectelly, we detected many genes in the common periphery of both genera, with overrepresented categories being conjugation, DNA modification etc. We propose that the common periphery could be inherited from the ancestor rather than completely arize from horizontal transfer. Some genes are universal in one genus but absent in the other, like many pathogenic genes for *Salmonella* and transporters for *Escherichia*. This may reflect

specialization of the genera. In the *Escherichia*/*Shigella* comparison such specialization genes were virtually absent, similar to the comparison of the commensal and pathogenic *Escherichia*. This provides an independent corroboration to the observation that shigellas belong to the *Escherichia* genus [3].

1. R.A.Welch et al. (2002) Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli, *Proc. Natl. Acad. Sci.*, **99(26):**17020-4.
2. H. Tettelin et al. (2005) Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial ''pan-genome'', *Proc. Natl. Acad. Sci.*, **102(39):**13950-5.
3. G.M. Pupo et al. (2000) Multiple independent origins of Shigella clones of Escherichia coli and convergent evolution of many of their characteristics, *Proc. Natl. Acad. Sci.*, **97(19):**10567-72.

Variance based identification of candidate genes using gene expression data

Ivan P. Gorlov[1], Jinyoung Byun[1], Hongya Zhao[2], Christopher Logothetis[1], and Olga Y. Gorlova[1]

[1]*The University of Texas MD Anderson Cancer Center, Houston, USA, ipgorlov@mdanderson.org*
[2]*Institute of Advanced Computing and Digital Engineering, Shenzhen, China*

Traditionally an identification of genes associated with cancer development is based on a comparison of mean expression values in cancerous (tumor) tissue and adjacent normal tissue[1,2]. We explored the possibility of using of the gene expression variance in prediction of cancer development. We focused on the two common types of cancer: prostate cancer (PCa) and lung cancer (LC).

We used KnowledgeNet bioinformatics tool to identify genes associated with lung cancer development (LC genes; n=200) and prostate cancer development (PCa genes; n=167). We found that LC genes have a higher interindividual variation of the expression level in normal lung tissue but not in normal prostate tissue (Figure 1), compared to all other genes. The similar observation was done for PCa genes. These results suggest that the *genes associated with cancer development tend to have a higher interindividual variance in normal tissue compared to the average gene in the human genome*.

In tumor tissue, the overall variance of the gene expression was higher than that in normal tissue: the average ratio of the variance of expression of a gene in the tumor to its variance of expression in the normal tissue for LC was 3.29±0.04, and for PCa it was 1.28±0.01. Both these

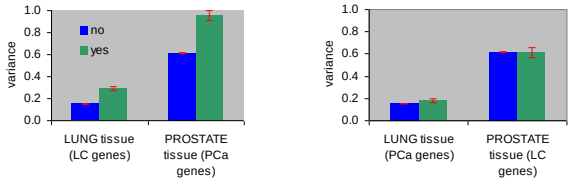


Figure 1. Interindividual variance in the expression level of cancer related genes in the same

values are significantly greater than 1, expected under the null hypothesis.

The increase in the variance in the transition from normal to tumorous tissue was more striking when cancer related genes were analyzed separately. For the LC genes the mean ratio of the tumor to normal lung tissue variances was 5.76±0.82 while for other genes it was 3.28±0.04. The difference is significant: t-test =3.03, N = 37690, P<0.0001. The difference is LC genes specific: for the PCa genes the mean ratio of the variance of expression in the lung cancer vs normal lung tissue is 4.21±0.56, which is not different from the other genes in the genome: t-test =1.66, N = 37690, P<0.21. Similar results were obtained for prostate cancer.

To explore if we can use variance-based approach to identify novel candidate genes we took the top 1% of the genes with highest increase in variance in tumor compared to the normal tissue, while showing no significant difference in the mean expression level. Note that those genes will be completely ignored by the traditional approach. We have estimated the correlation between the expression of these genes and Gleason score (GS) that is a key predictor of PCa progression [3]. We found that the mean absolute values of correlation coefficients (CC) was 0.14±0.01 for the genes with increased variance while the mean CC for other genes was 0.10±0.01 (the difference is significant P < 0.001). We used similar approach for lung cancer. We have identified 107 genes with increased variance in cancer but no significant difference between mean expression in normal versus cancer tissue. We found that those genes are more likely to be associated with differences in histologic types (adeno, squamos cell, and large cell carcinomas) compared to the other genes.

To summarize, our analysis demonstrates that: (i) Cancer associated genes (CAG) have a higher inderindividual variance in the corresponding normal tissue; (ii) CAGs also increase their variance in the transition from normal to tumor tissue more than other genes. An elevated variance of the cancer-related genes in tumor tissue might be a result of tumor heterogeneity. Different tumors may use different genes to progress. This may lead to increased variance of cancer associated genes in tumor samples. Our analysis also suggests that variance-based analysis allows identification of novel cancer related genes that are otherwise not detectable by standard approaches.

1. Lu, A.T., Salpeter, S.R., Reeve, A.E., Eschrich, S., Johnston, P.G., Barrier, A.J., Bertucci, F., Buckley, N.S., Salpeter, E.E. and Lin, A.Y. (2009) Gene expression profiles as predictors of poor outcomes in stage II colorectal cancer: A systematic review and meta-analysis. *Clin Colorectal Cancer*, **8:** 207-14.
2. Gorlov, I.P., Byun, J., Gorlova, O.Y., Aparicio, A.M., Efstathiou, E. and Logothetis, C.J. (2009) Candidate pathways and genes for prostate cancer: a meta-analysis of gene expression data. *BMC Med Genomics*, **2:** 48.
3. Molinie, V. (2008) [Gleason's score: update in 2008]. *Ann Pathol*, **28:** 350-3.

Derived SNP allele are more frequently used as a risk-associated variants in common human diseases

Olga Y. Gorlova, Jun Ying, Christopher I. Amos, Margaret Spitz, and Ivan P. Gorlov
The University of Texas MD Anderson Cancer Center, Houston, USA: oyorlov@mdanderson.org

Genome-wide association studies (GWAS) are a powerful tool to uncover the genetic architecture of complex human diseases [1]. The results of more than 200 GWAS have been published to date.[1] We used the GWAS data to address a question whether ancestral and derived (mutant) alleles are used as risk allele randomly, which is important for understanding evolutionary of the genetic control of common human diseases.

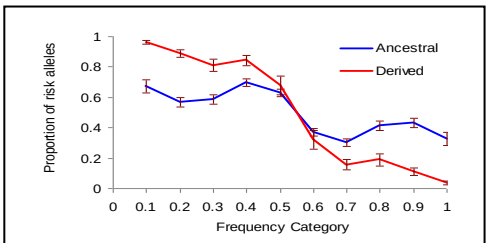


Figure 1. Proportions of the risk alleles in different allele frequency

Overall we found that rarer alleles are more likely to be "used" as a risk variant and common alleles are likely to be protective. Among derived alleles, the proportion of risk-associated variants for a given allele frequency was higher than that among ancestral alleles (0.96±0.01 vs. 0.67±0.04) (Figure 1). Among minor alleles (the left part of the distribution, Fig. 1), mean proportion of the risk variants was 0.84±0.05 for derived and 0.63±0.02 for ancestral alleles.

The analysis of early (typical onset of the disease before 30) versus late onset diseases (Figure 2) found that early onset diseases have a larger difference between ancestral and derived alleles in terms of probability of being a risk variant. Fore early onset diseases, the proportion of the risk variants among derived vs ancestral variants minor alleles was 0.79±0.06 vs 0.41±0.05, while the corresponding proportions for the late onset diseases were 0.78±0.07 vs 0.62±0.05. Therefore, human diseases that occur early in life are very likely to "use" a rare derived variant as a risk allele. For the late onset diseases risk alleles tend to be more uniform both in terms of population frequency and in terms of their evolutionary status, ancestral versus derived.

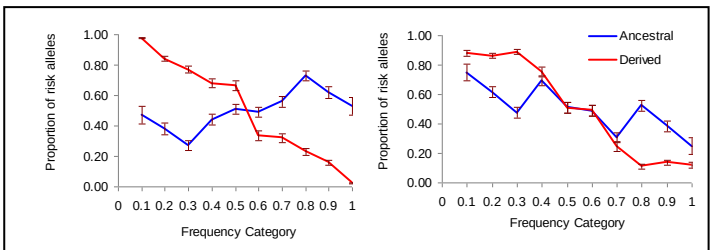


Figure 2. The proportions of risk variants in different frequency groups. Left panel – early onset diseases;

In brief we found that risk alleles are mostly derived low frequency (minor) variants. Both theoretical and experimental studies demonstrate that derived alleles generally have lower frequency [2,3]. The question is, therefore, whether the population frequency and derived status are independent predictors of the risk variant. Logistic regression analysis demonstrates that low frequency but not the derived status is a predictor of being the risk variant. The analysis of early onset diseases demonstrates, however, that derived status may be an independent predictor of the risk allele (data not shown). For the early onset human diseases a typical risk allele is a rare derived variant, while for the late onset diseases the usage of the allele as a risk variant is more random.

Mutations causing early onset diseases are likely subjected to negative selection. However such mutations can reach substantial population frequency because of the effects of the random factors like genetic drift, founder and bottle neck effects. For the late onset diseases, the negative selection does not affect allelic frequency. One can expect that mutations causing late onset diseases are evolutionary neutral, have stochastic dynamics and may completely replace ancestral allele leading to the situation when the risk allele is the ancestral (common) and protective allele is the derived (rare). This can explain why in the late onset diseases risk alleles are more random both in terms of population frequency and ancestral/derived status.

1. Johnson, A.D. and O'Donnell, C.J. (2009) An open access database of genome-wide association results. *BMC Med Genet*, **10**, 6.
2. Shastry, B.S. (2007) SNPs in disease gene mapping, medicinal drug development and evolution. *J Hum Genet*, **52**, 871-80.
3. Fredman, D., Sawyer, S.L., Stromqvist, L., Mottagui-Tabar, S., Kidd, K.K., Wahlestedt, C., Chanock, S.J. and Brookes, A.J. (2006) Nonsynonymous SNPs: validation characteristics, derived allele frequency patterns, and suggestive evidence for natural selection. *Hum Mutat*, **27:** 173-86.

De-Novo Discovery of Differentially Abundant DNA Binding Sites Including Their Positional Preference

Jens Keilwagen[1], Jan Grau[2], Ivan Paponov[3], Stefan Posch[2], Marc Strickert[4], Ivo Grosse[2]

[1]*Leibniz Institute of Plant Genetics and Crop Plant Research, Germany*
[2]*Martin Luther University, Germany*
[3]*University of Freiburg, Germany*
[4]*University of Siegen, Germany*
ivo.grosse@informatik.uni-halle.de

The development of new experimental techniques and the ensuing flood of large-scale genomics and epigenomics data has caused a renaissance of de-novo motif discovery. Two facts that complicate this task are that (i) many binding sites have a pronounced, but typically unknown, positional preference in their target regions and that (ii) they are weak and cannot be found from target regions alone but only by comparison with carefully selected control sets. Hence, we developed Dispom, a de-novo motif discovery program for learning differentially abundant motifs and their positional preferences simultaneously. Dispom outperforms existing programs based on benchmark data, succeeded in detecting a novel auxin-responsive element (ARE) in Arabidopsis thaliana substantially more auxin-specific than the canonical ARE, and turned out to be one of the top-scoring approaches in the latest DREAM challenge on the analysis of protein-binding microarrays.

MEASURING ALTERNATIVE SPLICING VARIABILITY

Roderic Guigo

Center for Genomic Regulation, Spain, roderic.guigo@crg.cat

We have developed statistical methodology to measure variation in gene expression and splicing ratios within and between populations, and to deconvolute the contribution of each of them to total variability in the abundances of individual transcripts.We have applied this methodology to estimates of transcript abundances obtained from RNA-seq experiments in lymphoblastoid cells from Caucasian and Yoruban individuals. We have found that protein coding genes exhibit reduced gene expression variability in human populations, and an even greater reduction in splicing ratios, with many genes exhibiting constant ratios across individuals. Consistent with this observation, we have found that genes involved in the regulation of splicing show less expression variability than human genes overall. While thereis correlation in splicing variability between populations, up to 10% of protein coding genes could exhibit population-specic splicing ratios. We estimate that about 50% of the total variability observed in the abundance of transcript forms can be explained by variability in transcription. A large fraction of the remaining variance can likely result from variability in splicing, although variability in splicing is uncommon without variability in transcription. Genes with high total variability (resulting from variability both in transcription and splicing) are particularly enriched in RNA binding functions. Consistent with this finding (and with the reduced variability of splicing factors), we have also found that long non coding RNAs show higher expression variability than protein coding genes. This suggests that variation in expression of long non coding RNAs may play an important role in establishing the molecular basis of intraspecies phenotypic individuality.

A combined approach to feature selection for multiclass microarray datasets

Georgy Gulbekyan, Valery Valyaev, Pavel Ivanov

M.V.Lomonosov Moscow State University, Russian Federation, gulbekyan@gmail.com
[2]MSU, Russian Federatio

The advent of expression microarray technologies made it possible to study transcriptome in various types of malignant cells [1,2]. It also allows one to diagnose several subtypes of the same tumor with relatively high precision to prescribe a proper administration [3]. Unfortunately, the usage of highthroughput microarrays for this purpose as well as Morphologic, Immunologic, Cytogenetic and Molecular biologic (MICM) classification is costly, time-consuming and don't yields perfect results. Therefore, the problem of detecting a small number of marker genes for reliable partitioning of different subtypes of the same pathology becomes extremely important.

To date, several algorithms addressing this problem have been proposed. The best classification accuracy on external validation set across all acute lymphoblastic leukemia (ALL) studies was reached using a well-known SVM-RFE (Support Vector Machine - Recursive Feature Elimination) approach [7].

We present a new approach to revealing marker genes in multiclass microarray datasets that combines and advances two well-proved approaches, namely, supervised classification and evolution simulation [5]. Since most of currently available cancer microarray datasets contain expression profiles for dozens of thousands of gene (in such a profile for a given gene, one expression value correspondes to a particular patient), a filtration step should precede further analysis. As most dequate to the problem of marker gene detection, we propose a method of expression profiles filtration in multiclass datasets that, first, approximates gene expression profiles in different classes by beta ditribution functions and, second, estimates gene relevance measure as a multiple convolution of such distributions.

We select the LIBSVM version of SVM technique to partition microarray samples into multiple classes using genes remained after filtration. We use a Leave-K-out Cross Validation to partition initial data into training and test datasets and to fit the parameters of SVM algorithm. After that, by randomizing initial data we estimate classification error for a large number of training/test partitions. Then, we generate mutations in a randomly chosen set of potential marker genes (predictor) from a filtered gene set and imitate simultaneous evolution of several such

predictors combined in a predictor pool. At each evolutionary epoch, we retain a predictor in or exclude it from the pool based on its classification power (quality measure). An elitism principle can also been added at this step. Also external validation using independent dataset was done.

We used dataset equal to [6] to compare our results to SVM-RFE approach. Our combined algorithm outperforms all previously ALL studies in terms of external validation accuracies. Results of applying the proposed algorithm to a model dataset as well as to results of experimental microarray tumor studies [3,5,6] will be presented.

1. Ramaswamy, S. and Golub, T.R. (2001) J. Clin. Oncol. 20: 1932-1941.
2. Segal, E. et al. (2002). Nat. Genet. 37: S38-S45.
3. Yeoh, E.J. et al. (2002). Cancer Cell 1: 133-143.
4. Jirapech-Umpai, T. and Aitken, S. (2005). BMC Bioinformatics 6: 148.
5. Ross, M. et al. (2003). Blood 102: 2951-2959.
6. Zhigang, Li et al (2009). Blood 114: 4486-4493
7. Guyon, I. et al (2002). Machine Learning. 46:389–422

Deep inside into invertebrate evolution: the molecular evolution modes of orthologous protein sequences

Konstantin Gunbin, Valentin Suslov, Dmitrij Afonnikov

Institute of Cytology and Genetics SB RAS, Russian Federation, genkvg@bionet.nsc.ru

The analysis of the molecular evolution modes of 2152 orthologous protein groups (OPGs) of invertebrates was made. Using MetaPhOrs [1] database we selected OPGs with at least one strictly confirmed orthologous sequence from each of the 10 invertebrate clades: (1) Cnidaria, (2) Tunicata, (3) Lophotrochozoa, (4) Caenorhabditis, (5) Low_chromadorea, (6) Arachnida & Crustacea (two polyphyletic taxa originated in Cambrian period), (7) Paraneoptera, (8) Lepidoptera & Coleoptera & Hymenoptera (paraphyletic taxa that arose prior to Permian-Triassic extinction), (9) Drosophilina и (10) Culicomorpha (taxa that arose after the Permian-Triassic extinction). Using the known tree topology (Figure 1) and calculated amino acid replacement matrices for each OPG the reconstruction of ancestral protein sequences by the PAML package [2] was made. They were used to calculate the observed number of each amino acid replacement types. To estimate the expected number of amino acid replacements, we conducted 1000 computer simulations of molecular evolution of each OPG using INDELible package [3], taking into account the all the peculiarities of the investigated OPGs. Comparison of

expected and observed changes of each replacement type was made using permutational test (10^5 permutations): we calculate the number of random samples, *M*, where the frequency of expected changes of a certain type was higher than the frequency of observed changes. The $M/10^5$ value assesses the occurrence probability *p* of a certain replacement type by chance.

It is of interest that the molecular evolution of insects and nematodes were characterized by significant increase in *R* value, the proportion of OPGs with atypical amino acid replacements ($p \leq 0.01$), in comparison with its mean value for Vertebrates (> 15%) and for taxa originated in Cambrian period (>8%). The absolute maximum of *R* value is typical for nematodes - a group with intensive lost of protein domain diversity [4].

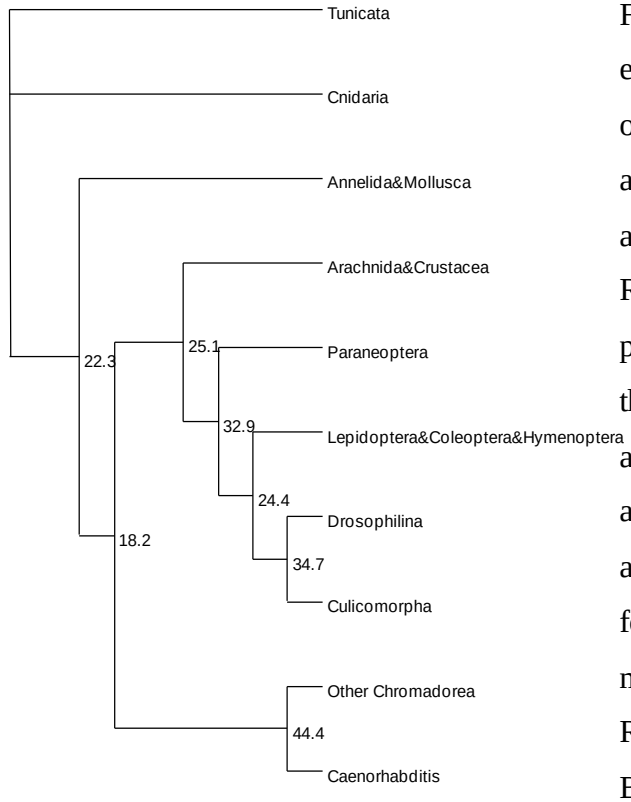


Fig. 1. Invertebrate phylogenetic tree. On each internal tree branch the proportion (%) of orthologous protein groups containing atypical, statistically rare ($p \leq 0.01$), types of amino acid replacements were shown.

Resolving the (8) and (6) clades of a phylogenetic tree leads to similar results, so the divergence of Insecta and Diptera accompanied by increasing *R* may be associated with the emergence of insects-angiosperms ecosystems [5] and with the formation of the characteristic Diptera morphology. The work was supported by RFBR grant No. 09-04-01641-a and Biosphere Origin and Evolution program.

1. L.P. Pryszcz et al. (2010) MetaPhOrs: orthology and paralogy predictions from multiple phylogenetic evidence using a consistency-based confidence score, *Nucl. Acids Res.*, doi: 10.1093/nar/gkq953.
2. Z. Yang (2007) PAML 4: phylogenetic analysis by maximum likelihood, Mol. Biol. Evol., **24**:1586-1591.
3. W. Fletcher, Z. Yang (2009) INDELible: a flexible simulator of biological sequence evolution, Mol. Biol. Evol., **26**:1879-1888.
4. C.M. Zmasek, A. Godzik (2011) Strong functional patterns in the evolution of eukaryotic genomes revealed by the reconstruction of ancestral protein domain repertoires, *Genome Biol.*, **12**:R4.
5. D. Grimaldi, M.S. Engel (2005) *Evolution of the insects*, NY: Cambridge University Press, 755p.

Human and Neanderthal miRNA genes are not so similar

Konstantin Gunbin, Dmitrij Afonnikov, Nikolay Kolchanov
Institute of Cytology and Genetics SB RAS, Russian Federation, genkvg@bionet.nsc.ru

In 2010, the first draft of Neanderthal genome was sequenced [1]. Of special interest are Neanderthal microRNAs (miRNAs), small RNAs that regulate gene expression by mRNA cleavage or repression of translation. Using human miRNA genome annotation based on MirBase, Richard E. Green et al. were found only one miRNA possessed a fixed substitution (hsa-mir-1304) and one case of a fixed single nucleotide insertion (AC109351.3) [1]. We used human miRNA genome annotation based on Ensembl rel. 61 and UCSC Neanderthal Genome Analysis Consortium Tracks for analysis of possible ways of the Neanderthal miRNA evolution. All short reads partially encompassing human pre-miRNA genome regions were excluded from our analysis. Thus, final sample comprised 187 human-Neanderthal miRNA alignments (~10% from all human miRNAs annotated in Ensembl rel. 61). In the further analysis we excluded all (single and multiple substitutions) cases with C to T and G to A nucleotide substitutions between human and Neanderthal, which are caused by deaminated residues in Neanderthal fossils [1] and also excluded cases with single nucleotide substitutions of other types with significant probability of its fossil nature. As a result we found 11 (~6% of our sample) Neanderthal miRNAs with multiple substitutions or deletions. Using Ensembl comparative data, we showed that at least 8 of them (Table 1) possessed unevenly distributed nucleotide substitutions or deletions, which are not characteristic of human and other primates (*Pan troglodytes, Gorilla gorilla, Pongo pygmaeus, Macaca mulatta* and *Callithrix jacchus*).

Table 1. Neandertal miRNAs with severe changes from human and primates.

Ensembl ID	Changes from human ortholog (positions on the basis of human miRNAs)
ENSG00000221598	A10G; A19G; G13T
ENSG00000221378	T32C; T63C; T70C; T83C
ENSG00000220996	Deletion(48-50)
ENSG00000211520	T3G; T11G; A21G
ENSG00000208036	C73G; C76A
ENSG00000207758	A79G; A82G
ENSG00000207728	C44G; A47G; A52G
ENSG00000207579	C11G; G13C; C14A

It is likely that these miRNAs may reflect evolutionary changes in Neanderthal genome related to its adaptation to severe environment of ice ages or depopulation and extinction just after ice ages. The work was supported by RFBR grant No. 09-04-01641-a and Biosphere Origin and Evolution program.

1. R.E. Green et al. (2010) A draft sequence of the Neandertal genome, *Science*, **328**:710-722.

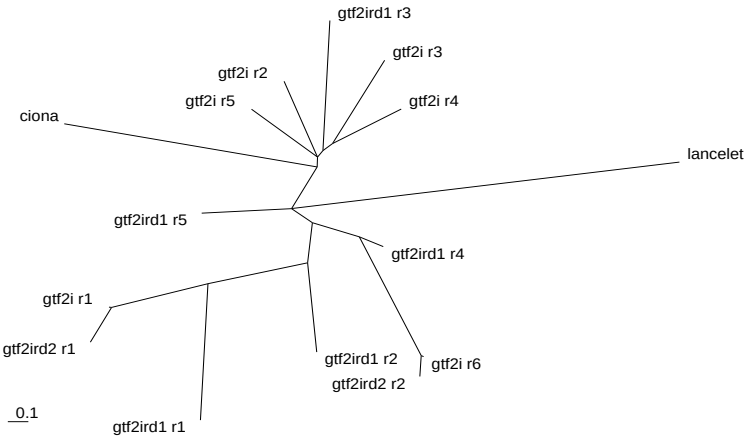
GTF2I DOMAIN: STRUCTURE, EVOLUTION AND FUNCTION

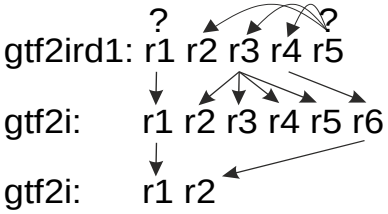
Irina Medvedeva[1], Konstantin Gunbin[1], Vladimir Ivanisenko[1], Anatoly Ruvinsky[2]

[1] *Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia, genkvg@bionet.nsc.ru*
[2]*The Institute for Genetics and Bioinformatics, University of New England, Armidale, NSW 2351, Australia*

The GTF2I gene family consists of *gtf2i*, *gtf2ird1* and *gtf2ird2* genes encoding transcriptional factors and the first two of them involved in Williams-Beuren syndrome if mutated [1]. The main characteristic of this gene family is the presence of several so called GTF2I repeats. There are 5 such repeats in *GTF2IRD1 (R1-R5)*, 6 in *GTF2I (R1-R6)* and 2 in *GTF2IRD2 (R1-R2)*. The investigation of the exon-intron structure evolution of these genes and the study of the tertiary structure differences of the GTF2I domains were our aims.

We collected the GTF2I repeats from about 20 species from *Ciona intestinalis* to *Homo sapiens* using the information about gtf2i-like genes from databases and retrieving information from genomes, assemblies and EST collections. Then we reconstructed the phylogenetic tree using Bayesian and maximum likelihood methods, and found that R5 (*gtf2ird1*) is the most close repeat to the homologous sequences in squirt and lancelet, which represents most probable ancestors of GTF2I repeats.





We applied the comparative analysis identified the Activation Domain which is essential for transcription activation. It's located between R1 and R2 (*gtf2i* and *gtf2ird2*) and between R2 and R3 (*gtf2ird1*) that confirms the common ancestor of GTF2I repeat for R2 (gtf2i) and R3 (gtf2ird1). Based on our hierarchical classification of GTF2I repeats and recent research [2] we conclude that GTF2I gene family have raised from the two genome duplication events early in vertebrate evolution.

Comparison of molecular evolution model of the GTF2I family with real data using method implemented in SAMEM [3] discovered several types of statistically rare aminoacids substitutions which characterized by greatest physicochemical changes. We used the program package PDB3DScan for structural alignment to find the differences in the 3D structure of GTF2I repeats in gtf2i gene and identified that those statistically rare substitution mostly situated in loop regions or in the beginning or in the end of helices. We showed that the helix in the GTF2I structure in the first repeats of gtf2i, gtf2ird1, gtf2ird2 is much less conservative then in the other repeats. We also found the unique mutational pattern for each GTF2I repeat that could be caused by their functional specialization (for example, DNA-binding [4]). The work was supported by RFBR grant No. 09-04-01641-a and Biosphere Origin and Evolution program.

1. Bayarsaihan, D. et al. (2002) Genomic organization of the genes Gtf2ird1, Gtf2i, and Ncf1 at the mouse chromosome 5 region syntenic to the human chromosome 7q11.23 Williams syndrome critical region. *Genomics* **79**(1):137-43.
2. P. Dehal, J.L Boore. (2005) Two Rounds of Whole Genome Duplication in the Ancestral Vertebrate, *PLoS Biol* **3**: e314.
3. K.V. Gunbin et al. (2010) A computer system for the analysis of molecular evolution modes of protein-encoding genes (SAMEM), *Moscow University Biological Sciences Bulletin*, **65**: 142-144.
4. S.J. Palmer et al. (2010) Negative autoregulation of GTF2IRD1 in Williams-Beuren syndrome via a novel DNA binding mechanism, *J Biol Chem.* **285**:4715-4724.

Supramolecular Complexes of the *A. tumefaciens* Virulence Protein VirE2

Yuriy Gusev, Irina Volokhina, Mikhail Chumakov

Institute of Biochemistry and Physiology of Plants and Microorganisms, Russian Academy of Sciences, 13 Prospekt Entuziastov, Saratov 410049, Russia, yuran1989@yandex.ru, chumakov@ibppm.sgu.ru

Soil bacteria of the genus *Agrobacterium* are a natural vector for the transfer of genetic information (T-DNA) into the plant cell. The T-DNA nonspecifically integrates into the host's chromosome and is inherited at subsequent cell divisions. It is believed that VirE2 protein forms a membrane-spanning pore or a channel for the promotion of translocation of short nucleotides across the membrane; this pore may form from four VirE2 subunits [1, 2]. The VirE2 protein then nonspecifically and cooperatively attaches to ssDNA, forming a "T-complex" in the host cytoplasm [3].

The aims of this work were to study the interactions between VirE2 proteins and to perform a computer simulation of VirE2's pore-forming capacity. For formation of a complex consisting of two and four VirE2 proteins by using the 3D model described by Dym et al. [4],

we used the programs Hex (http://hexserver.loria.fr) and GRAMM-X (http://vakser.bioinformatics.ku.edu/resources/gramm/grammx). The structure composed of two VirE2 proteins, presented in figure 1a, had a size of 8.4x7.3x5.4 nm. A channel with a diameter of 2.5 nm in a complex formed from two VirE2 proteins might form between Asp457 of both VirE2 proteins. However, in the bore of this channel were exposed the ends of a motile interdomain loop (amino acid residues Glu470 of both VirE2 proteins), which narrowed the channel to 0.9 nm. The model structure built from four VirE2 proteins, presented in figure 1b, had a channel of 3.8 nm in size (measured between Lys488 VirE2 protein (number 1) and Lys488 VirE2 protein (number 3)). The complex from four VirE2 proteins had a size of 15.3x11x6.4 nm. Using the dynamic light scattering method, we evaluated *in vitro* the hydrodynamic diameter of VirE2 particles in aqueous solutions by photon correlation spectroscopy by using a Zetasizer Nano instrument (model ZEN 3500; Malvern Instrument LTD, UK). Assessments were made by the change in the initial hydrodynamic diameter of molecules in solution before and after coincubation. The size of recombinant VirE2 from *E. coli*, after chromatographic purification on an Ni-NTA-agarose column, was 12 nm (51% of all particles in solution); 47% of all particles were represented by 115-nm aggregates; and only a small portion (2%) contained 2,800-nm particles.

In summary, we have shown, by using computer modeling, the possibility of pore formation in supramolecular complexes formed from two and four VirE2 proteins with channel sizes of 2.2 and 3.7 nm, respectively.

We thank B.N. Khlebtshov for his help with light-scattering experiments.

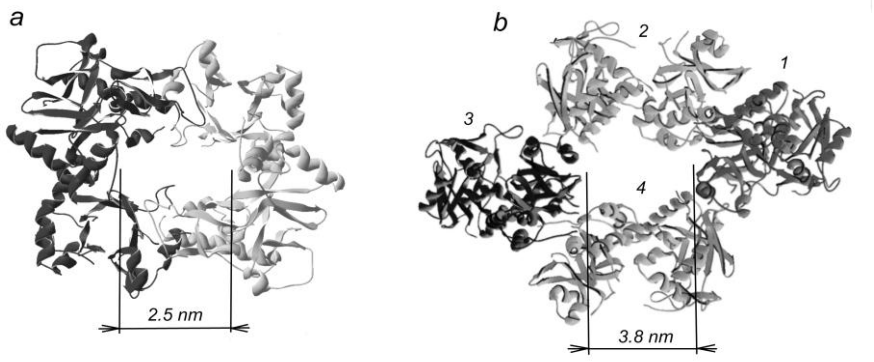


Fig. 1. *Complexes consisting of a) two and b) four VirE2 proteins using the model described by Dym et al.[4].*

1. F. Dumas et al. (2001) An *Agrobacterium* VirE2 channel for transferred-DNA transport into plant cells, *Proc. Natl. Acad. Sci. USA*, **98**:485–490.
2. M. I. Chumakov et al. (2010) Study of the ability of agrobacterial protein VirE2 to form pores in membranes, *Biochemistry (Moscow). Supplement. Series A. Membrane and Cell Biology*, **27**:449–454.
3. S.B. Gelvin (1998) *Agrobacterium* VirE2 protein can form a complex with T strands in the plant cytoplasm, *J. Bacteriol*, **181**:4300–4302.
4. O. Dym et al. (2008) Crystal structure of the *Agrobacterium* virulence complex VirE1-VirE2 reveals a flexible protein that can accommodate different partners, *Proc. Natl. Acad. Sci. USA*, **105**:11170–11175.

Visualization and Analysis of a Cardio Vascular Disease-related Biological Network combining Text Mining and Data Warehouse Approaches

Ralf Hofestädt[1], Björn Sommer[1], Evgeny Tiys[2], Benjamin Kormeier[1], Klaus Hippe[1], Sebastian Janowski[1], Timofey Ivanisenko[2], Anatoly Bragin[2], Patrizio Arrigo[3], Pavel Demenkov[4], Alexey Kochetov[2], Vladimir Ivanisenko[2], Nikolay Kolchanov[2]

[1] *Bielefeld University, Germany, ralf.hofestaedt@uni-bielefeld.de*
[2] *Institute of Cytology and Genetics SB RAS, Russian Federation*
[3] *CNR ISMAC, Italy*
[4] *Sobolev Institute of Mathematics SB RAS, Russian Federation*

Detailed investigation of socially important diseases with modern experimental methods has resulted in the generation of large volume of valuable data. However, analysis and interpretation of this data needs application of efficient computational techniques and systems biology approaches. In particular, the techniques allowing the reconstruction of associative networks of various biological objects and events can be useful. In this publication, the combination of different techniques to create such a network associated with an abstract cell environment is discussed in order to gain insights into the functional as well as spatial interrelationships. It is shown that experimentally gained knowledge enriched with data warehouse content and text mining data can be used for the reconstruction and localization of a cardiovascular disease developing network beginning with MUPP1/MPDZ (multi-PDZ domain protein).

A systems biology approach to unravel the underlying functional modules involved in autism

Roser Corominas[1], Shuli Kang[1], Guan Ning Lin[1], Xinping Yang[2], Yun Shen[2] , Pascal Braun[2], Jonathan Sebat[1], David E. Hill[2], Kourosh Salehi-Ashtiani[2] , Marc Vidal[2], Tong Hao[2] and Lilia M. Iakoucheva[1]

[1] *Department of Psychiatry, University of California, San Diego, La Jolla, CA 92093*
[2] *Center for Cancer Systems Biology (CCSB), Department of Cancer Biology, Dana-Faber Cancer Institute, Boston, MA 02115*
*Corresponding authors: Lilia M. Iakoucheva (*lilyak@ucsd.edu*) and Tong Hao (*tong_hao@dfci.harvard.edu*)*

Autism is a neurodevelopmental disorder with strong genetic basis. Recent genetic studies using families and unrelated autism case control samples have firmly established that rare copy number variants (CNVs) play significant role in determining risk for autism. Spontaneous (*de novo*) CNVs, in particular, have been implicated as major causal genetic risk factors for sporadic form of autism. The number of genes that have now been firmly established as strong candidates for autism is large, and these genes are functionally heterogeneous. Hence, it is important to understand how these multiple genes and their protein products interact within the context of cellular pathways. Here, we investigate autism from the systems biology perspective. We use an integrated experimental genetics and functional protein-protein interaction network-based approach to identify key pathways/networks/functional modules that are involved in autism. We hypothesize that perturbations of these pathways might lead to manifestation of broad disease phenotypes. Our approach consists of three steps: (1) perform a large-scale discovery of alternatively spliced isoforms of autism gene candidates using our recently developed high-throughput isoform discovery pipeline that incorporates parallel 454 FLX sequencing and computational analysis platforms; (2) identify and clone mutant transcripts of the genes disrupted by the breakpoints of genomic deletions and duplications in autistic patients; (3) build an interactome of autism candidate genes, their alternatively spliced variants and mutant transcripts to define key functional modules involved in ASD. For step 1, we attempted to systematically clone full-length open reading frames (ORFs) and the splice isoforms corresponding to 191 autism candidate genes from fetal and adult human brain RNA. Next generation 454 sequencing, assembly of the reads and further informatics analyses indicated that we were able to clone one or more ORFs/isoforms from 130 genes (69% success rate). After redundancy removal, we were able to capture 365 distinct isoforms encoded by these genes (~2.8 isoforms per gene on average). When compared to isoforms recorded in several major databases, 207 out of 365 isoforms (56.7%) appear to be novel. Significantly more (~1.3 fold, p=0.014)

novel isoforms, and isoforms utilizing non-canonical splicing sites (~1.6 fold, p=0.002), have been identified from fetal brain than from adult brain, possibly suggesting a more complicated regulation of pre-mRNA during brain development. We are rapidly expanding the knowledge of the isoform space of autism candidate genes. For step 2, we have identified 35 de novo CNVs in autism patients that could potentially produce aberrant mRNA transcripts/proteins. Some interesting examples include an in-frame deletion of several exons of DNM3 gene leading to a shorter truncated protein, and deletions producing a reading frame shift in NRXN1 gene that is repeatedly recognized as a risk factor for autism. We are currently attempting to clone the disrupted genes from the patients' lymphoblastoid cell lines. For step 3, using high-throughput yeast two-hybrid system (Y2H), we have screened 166 autism candidate genes against human ORFeome consisting of 12757 unique ORFs. We were able to detect 430 interaction partners for 115/166 autism genes, with the total of 1108 interactions between them. Currently, these interactions are being retested and validated; therefore here we are presenting preliminary results for the unretested dataset. Amongst 1108 interactions, 82 unique interactions (7.4%) have previously been observed in at least one other interaction database, whereas 1026 (92.6%) appear to be novel. Similarly to other interaction networks, autism network is characterized by the presence of a small number of highly connected proteins (~9%) and a large number of proteins interacting with just one or a few partners suggesting its scale-free topology. The shortest path length for autism network (3.21) is significantly shorter than for the latest human interactome HI2 (3.86, Wilcoxon $p<0.004$). To better understand the relationships among autism network genes, we first ranked them based on their ability to directly interact with partners from autism network than from 1000 randomly rewired networks with the same number of nodes and edges. Next, we ranked the interacting partners (preys) of autism network genes (baits) by calculating the significance of their connectivity in autism network compared to their connectivity in the human interactome HI2. Fifty preys passed the Chi-square significance test with p-value cutoff of 0.0002 with Bonferroni correction. Among them, several interesting targets are already emerging such as LNX2 (Notch signaling), CTBP2 (Wnt signaling), MEOX2 (transcription factor) and others. Our future work is focused on retesting autism interactome, building interaction networks of splice variants and mutant transcripts from the patients, and on developing statistical methods for detecting new candidate genes and functional modules involved in autism spectrum disorders.

Acknowledgements: This work is supported by the National Institute of Health grants ARRA R01HD065288 and R01MH091350, Iakoucheva and Hao (PIs).

Inter-SNP distances and SNP distribution in the human genome

Elena Ignatieva, Victor Levitsky, Nikolay Yudin

Institute of Cytology and Genetics, Novosibirsk, Russian Federation,
eignat@bionet.nsc.ru , levitsky@bionet.nsc.ru, yudin@bionet.nsc.ru

The results of the pilot phase of the 1000 Genomes Project were approximately 15 million SNPs, 55% of which were previously undescribed [1]. Mutation frequencies are known to vary along a nucleotide sequence. Nucleotide positions with an exceptionally high mutation frequency are called hotspots. The examination of mutation hotspots provides evidence that they arise due to some structural features of hotspot subsequences. Accordingly, the local DNA sequence contexts of hotspots (mutable motifs) were determined [2]. However, genome-scale patterns of SNP distribution have not been systematically investigated. Specifically, the low-coverage data from 1000 Genomes Project allowed to address a long-standing debate about whether recombination has any local mutagenic effect. It was concluded that, although recombination may influence the fate of new mutations, for example through biased gene conversion, there is no evidence that it influences the rate at which new SNP variants appear. A two to four-fold increase was found in SNP frequency in the human genome at positions immediately adjacent to the boundaries of mononucleotide repeats, relative to that at more distant bases [3]. Study of the inter-SNP distance is of great importance for experimental researches. Sensitivity and specificity of most experimental methods of SNP typing are strongly decreased from the presence of additional SNP variants in the neighborhood with the target SNP. One of the most popular tools for screening of pathogenic SNP variants, high-resolution melting curve analysis (HRM) is especially vulnerable, because for SNP screening, DNA fragments of 150–250 bp are usually used [4]. In this study, we analyzed the pattern of SNP distribution among the different regions of human genome using data from the 1000 Genomes Project which were extracted from the dbSNP131. We also studied the DNA context features and functional characteristics of nucleotide sequences with adjacent SNPs as well.

It was found that 1000 genome project SNPs are localized non-uniformly among human genome. Contrary to expected (stochastic) SNP distribution (one SNP occurs on average every 268 bp), more than half SNPs (69%) occurred at distances less than 250 bp. About 1.3% of SNPs occur in neighboring positions (adjacent SNPs). It is 3 times more than expected accident frequency and 1.5 times more then the observed number of SNPs separated by one nucleotide. SNP density was dependent on SNPs localization across different parts of gene. Low SNP density was found in the vicinity of transcription start sites, 5’ and 3’ ends of introns. This

observation is in good agreement with important functional roles of these regions. It was found that SNP distribution in exons has the periodicity divisible by 3. This reflects the fact that according to the standard genetic code many synonymous substitutions may occur at the third codon positions. We also investigated the DNA context features of nucleotide stretches with adjacent SNPs. There were several general DNA context features of the stretches in different regions of genes (exons, 5'UTRs, 3'UTRs): the frequency of CpG was found to be higher among the central positions and AT-content was found to be higher in positions adjacent to the boundary of the stretch. We concluded that the possibility of existence of additional SNP variants in melting fragment requires careful consideration for improvement of specificity of new HRM assays. For example, if the melt profile contains more than one domain it will be warranted to resequence of this DNA sample.

Acknowledgements This research was supported by the Ministry of Education and Science of RF (contr. № P721), SB RAS (Project № 119) and RAS (Programs №№ B.25, B.27, A.II.6)

1. 1000 Genomes Project Consortium et al. (2010) A map of human genome variation from population-scale sequencing, Nature,467:1061-1073.
2.Rogozin I.B. et al. (2003) Computational analysis of mutation spectra, Brief. Bioinform. 4:210-27.
3.Siddle K.J. et al. (2011) Bases adjacent to mononucleotide repeats show an increased single nucleotide polymorphism frequency in the human genome, Bioinformatics, 27:895-898
4.Vossen R.H. et al. (2009) High-resolution melting analysis (HRMA): more than just sequence variant screening, Hum. Mutat., 30:860-866.

New Clustered Regularly Interspaced Short Palindrome Repeats in Xanthomonads

Alexander Ignatov[1], Dinara Mallabaeva[1], Doug Luster[2], Norman Schaad[2]

[1]*Center Bioengineering RAS, Russian Federation, an.ignatov@gmail.com*
[2]*FDWSRU-ARS, USDA, Fort Detrick, MD, USA, United States*

Analyses *in silico* carried out in the region of simple repeated DNA revealed the presence of conservative DNA fragments with 34 bp terminal inverted repeats (TIRs) in all complete genomes of xanthomonads. From the alignment of DNA regions from 15 complete genomes of xanthomonads we observed the fragments to be miniature mobile elements with consensus sequence GTA GGV GCG CVC BNG NGC GCG ANG NVG NNB BNN CGN NVV NNC BNC NTC GCG CNC NVG BGC GCB CCT AC (68bp). The lengths of XAMIS were from 67 to 69 bp; all folded into palindrome structures. Whole genome sequences of xanthomonads were analyzed for the Xanthomonads Miniature Insertion Sequences (XAMIS) repeats. From 11 to 136 copies of XAMIS were found in these genomes (Table 1). All but a few XAMIS were located in intergenic regions and were at least 13 bp from nearest gene sequence. Most were found in several copies (2-6) in the same intergenic region in *Xcc* genomes (Fig. 1). Whereas other xanthomonads had only one to two XAMIS copies in one site. Analysis of XAMIS distributions revealed they often existed near virulence, membrane biosynthesis, and transporter genes. Average genetic difference between different XAMIS ranged from 0.189 to 0.386 within a single strain, and in the same range – between different species. The genomic distribution of XAMIS showed high similarity between strains of the same species. Repeats with partial similarity to XAMIS were found in genomes of *Stenotrophomonas* sp., *Ralstonia* sp., and *Pseudomonas* sp. but they were not inverted repeats. We evaluated possible function of XAMIS as new CRISPRs (Clustered Regularly Interspaced Short Palindrome Repeats) elements, involved in bacteriophage resistance of the bacteria (Pourcel, et al. 2005), and found some evidence of presence of phage fragments in XAMIS-flanked repeats (Fig. 1).

Acknowledgements: This work was supported by ISTC projects 3431 and 3978

Reference: C. Pourcel, et al. (2005). «CRISPR elements in *Yersinia pestis* acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies». *Microbiology* **151**: 653-659.

CCTTGAGGAGGCGGGGATTGGGGAGTTGGGAATCGGGATCGTAAAAGCGGTGGCTGCGGCCTGATTGGTA
phage lambda, phiL7
ACTGCGCCTGATCGGTGGGTGTCGCCAGTTGCATGCCCGAGTACCAGACCGCAATGACGTCACGCTTGTGTAGGAGCGCGCTTGCG
CGCGATGAGGCGTTACCGGGAGAGCTTCATCGCGCG
phage CMP1, kappa, K139, V86
CAAGCGCGCTCCTACGCGGCAGGTCTTCCGGCCTGATGCCTTCGACTGCAAGGCTTGCCCACGCACGCGCGTCGGAGCAGCAACCA
TCCAGAACGCAGCGACATCACGCTTGTGTAGGAATGCGCTTGCGCGCGATGGGCGTTATCGGGAGAGCTTCATCGCGCGCGAGCGC
GCTCCTACGAAACAGGCCTTCCGGCCTGGATGCGTTCGACTGCAAGGCTTGCCCACGCACGCACGCACGCGGGTCGGAGCAGAACC
ATACAGACCGCAGTGACATCACGCTTGTGTAGGAGCGCGCTTGCGCGCGATGAGGCGTTACCGGGAGAGCTTCATCGCGCGCGAGC
GCGCTC
Molluscum contagiosum virus
CTACGCGGCATGTCTTCCGGCCTGATGCCTTCGACTGCCGGGCTTTCCCATGCACGCACGCGCGTCGGAGCAACAACACAGACCGC
AGGTGACATCACGCTTGTGTAGGAGCGCGCTTGCGCGCGATGGGGCGTTATCGGGAGAGCTTCATCGCGCGCGAGCGCGCTCCTAC
GAGACGGGTCTTCCGGCCTGAGGCCTTCGATTGCCGGCTTGCCCACGCACGCACGCGCATCAGGGCAACGACCATCCAAACCGCAG
CGACATCACGCTTGTGTAGGAGCGCGCTTGCGCGCGATGGGGCGTTACCGTGAGGACTTCATCG
Mycobacterium phage Pacc40, PMC, Llij, virus RRV
CACGCGAGCGCGCTCCTACGCGGCAGTTCTTCCGGCCTGATGCCTTCGACTGCCGGGCTTTCCCACGACGCACGCGCGCTCGGGGC
AACAACCATCCAGACTGCAGTGACCTGCGCTCCTACGCGGCGTTTGTGGCGGTCTTCGTGGATGCTCGGCATGCTTGCCACATCAC
CCAACTAGCCCCCCCCCCCCAAAGGCGTCGCGCCGCTGCACAGCCTGCATGACCCAGCGCCAGGCCACGCTTCCA

Figure 1. Sequence of intergenic region 4355510-4356595 (1085 bp) region of strain X.campestris pv. campestris ATCC 33913 containing 5.5 repeats of long XAMISs (**bold underlined**) and virus-like sequences (*italic underlined*)

Table 1. Number of XAMISs in several genomes of xanthomonads.

Strain	GenBank accession	Complete XAMISs copies
Xanthomonas campestris pv. *campestris* ATCC 33913	NC_003902.1	115
X. campestris pv. *campestris* 8004	NC_007086.1	98
X. campestris pv. *vesicatoria* 85-10	NC_007508.1	48
X. axonopodis pv. *citri* 306	NC_003919.1	44
X. oryzae pv. *oryzae* KACC10331	NC_006834.1	35
X. oryzae pv. *oryzicola* BLS256	AAQN00000000.1	33
Stenotrophomonas maltophilia K279a	AM743169	11

Chromatin folding in eucaryotes: matching 3D genome structure to polymer models using molecular dynamics simulations.

Maxim Imakaev, Leonid Mirny

Massachusetts Institute of Technology, 77 Massachusetts ave E25-524, Cambridge, MA, 02139, USA
imakaev@mit.edu, leonid@mit.edu

Several genomic and optical techniques for probing a high-order structure of genome have been recently developed.[1,3] These techniques however do not provide comprehensive 3D picture of genome folding. Optical methods allow tracking only a few loci, while the chromosome conformational capture methods give a contact probability map, rather than spatial coordinates of loci. Here we use GPU assisted Molecular Dynamics[4] simulations to infer, characterize and visualize chromatin folding using experimentally obtained contact maps. In particular we focus on the underlying polymer structure as seen by scalings, on domain formation and chromatin interactions.

A simplest quantitative description of contact map of a polymer is scaling of a contact probability with genomic distance. A recent study[1] suggested that human DNA contact probability scales inversely with the genomic distance, and thus it is consistent with a fractal globule (FG)[2] model as opposed to the equilibrium globule. It was generally believed that the fractal globule is a long-lived state and is stabilized by by topological interactions, i.e. the fact that the polymer chain cannot cross itself.[2]

Using MD simulations we model mitotic chromosome decondensation subject to specific intrachromosomal interactions. We observe that the contact probability scaling gradually changes with time and reaches that of the equilibrium globule. We also study the diffusion of the FG and see that it reaches the equilibrium-like state with the same speed, in time t~N^(4/3), that is much faster than proposed t~N^3 time for true equilibration of a polymer chain.[2]

We perform a chromosome-by-chromosome analysis of the contact probability in the Hi-C data and show that small chromosomes look more like partially equilibrated globules, while larger chromosomes show similarity to the mitotic-like chromosome packing. This suggests that experimental data can be better described by a non-equilibrium state of a chromatin fiber after decondensation, than by the FG.

Another property of the 3D chromatin structure is compartmentalization of open and closed chromatin. Confirmed both by Hi-C, 5C, FISH and electron microscopy, it shows that there are two chromatin domains that are spatially segregated: open, active and highly-express chromatin, and closed (repressed). We show that weak attraction of only one type of chromatin to itself is sufficient to spatially separate two domains without affecting diffusive properties of

the chromatin. As a model for this, we simulate attraction of all highly-expressed genes on a single chromosome, and get a good agreement with a contact map obtained by Hi-C methods. We also check what number of specific interactions is sufficient to successfully form these two compartments, and get an estimate of hundreds of interactions per small chromosome.

1. E. Lieberman – Aiden, N. Van Berkum (2009), Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome, Science, **326**:289-293
2. A. Yu. Grosberg, S.K. Nechaev (1988), The role of topological constraints in the kinetics of collapse of macromolecules, Journal de Physique **49:** 2095-2100
3. N. van Berkum, J. Dekker (2009), Determining Spatial Chromatin Organization of Large Genomic Regions Using 5C Technology, Methods in Molecular Biology, **567**: 189-213
4. M. S. Friedrichs, P. Eastman (2009), Accelerating Molecular Dynamic Simulation on Graphics Processing Units., J. Comp. Chem., **30(6)**:864-872

Analysis of cleaved N-terminal sequences coming from MS/MS proteomics for E.coli and S.cerevisiae

Dmitry Ivankov[1], Stefano Bonissone[2], Pavel Pevzner[2], Dmitrij Frishman[3]

[1]*Technische Universitaet Muenchen, Germany, ivankov13@gmail.com*
[2]*University of California San Diego, United States*
[3]*Technische Universitaet Muenchen, German*

Sub-cellular localization is an important aspect of protein function. It determines the molecular environment in which proteins operate, in particular the availability of interaction partners. In bacteria, the majority of proteins execute their function in the cytoplasm, but a sizeable fraction of gene products are directed to other cellular locations – the cytoplasmic membrane, cell wall and extracellular space in Gram-positive bacteria and the cytoplasmic membrane, the periplasm, the outer membrane and the extracellular space in Gram-negative bacteria [1]. In eukaryotes more than 20 different localizations can be distinguished [2] with the major ones being cytoplasm, nucleus, mitochondria, extracellular space, various kinds of membranes, as well as chloroplasts in plants.

Post-translational transport of proteins to cellular compartments involves translocation across at least one membrane. In many cases proteins are targeted to their cellular destinations by means of short sequence motifs that are involved in molecular interactions with membrane receptors. In particular, proteins directed to the secretory pathway, to mitochondria, and to chloroplasts typically contain signal peptides, mitochondrial targeting peptides, and chloroplast transit peptides, respectively, located at their N-terminus. These three types of targeting motifs that are cleaved off after translocation by signal peptidases constitute the focus of our research project.

The two major pathways for exporting unfolded and folded proteins are the essential and universal Sec (general secretion) system and the Tat (twin arginine translocation) system [3]. The Sec translocase is found in the cytoplasmic membrane of all bacteria, archaea, in the thylakoid membrane of plant chloroplasts, and in the endoplasmic reticulum of eukaryotic cells. In eukaryotes proteins secreted through the Sec-pathway are further directed to other compartments via the vesicle sorting route. The Tat pathway exists in many bacteria and archaea, in the thylakoid membranes of plant plastids, and, presumably, in plant mitochondria [4].

Here we examine targeting signals of E.coli and S.cerevisiae derived by correlating mass spectral data with the genome structure in the framework of a proteogenomics approach. The analysis focuses exclusively on peptides with a non-tryptic N-terminus with no upstream coverage [5]. Such peptides indicate that the upstream N-terminal protein fragments, not directly observed in MS experiment, were cleaved in vivo by some peptidase. Analysis of the E.coli and S.cerevisiae genomes as well as of the previously studied S.oneidensis [5] genome shows that in most cases it is the signal peptidase I and mitochondrial peptidases (in case of yeast).

Analysis of E.coli shows that out of 42 cleaved N-terminal peptides detected by MS proteomics about 60% are signal peptides, about 30% are not signal peptides, and the remaining ~10% can not be easily classified. Out of signal peptides 80% comes from periplasm proteins, while 12% comes from inner/outer membrane proteins. 8% are exported via the Tat-signal pathway.

In S.cerevisiae the number of cleaved N-terminal peptides detected in MS proteomics is an order higher than in E.coli, namely, 413 peptides. Of them more than one third are confident signal/transit peptides and about 40% are probably not signal or transit peptides. About 70% of the confident signaling sequences are mitochondrial transit peptides, while the remaining sequences are apparently cleaved by signal peptidase I. Comparison of the found mitochondrial target peptides with those annotated in UniProt reveals different scenarios of action of mitochondrial peptidases: 1) the so called ‘R-2’ motif is indicative of the cleavage by the mitochondrial processing peptidase (MPP), 2) the ‘R-3’ motif corresponds to processing by MPP followed by intermediate cleavage peptidase (Icp55), 3) the ‘R-10’ motif mediates the action of the MPP followed by octapeptidase (Oct1), and finally 4) the ‘R-11’ may be the evidence of the experimentally yet unknown combined action of MPP, Oct1, and Icp55.

To summarize, proteogenomics data for E.coli and S.cerevisiae shows that most of the in vivo proteolytic cleavages in the cell detected by MS are cleavages of signal and transit peptides. Even for such well-studied model organisms we observed many signal/transit peptides not yet confirmed by any experiment. Extrapolation of the results to the poorly studied genomes implies that the proteogenomics approach has a great potential for expanding our knowledge about signal sequences.

1. J.L.Gardy, F.S.L.Brinkman (2006) Methods for predicting bacterial protein subcellular localization. Nat Rev Micro, 4:741-751.
2. A.Pierleoni, P.L.Martelli, P.Fariselli, R.Casadio (2007) eSLDB: eukaryotic subcellular localization database. Nucleic Acids Res, 35:D208-212.
3. P.Natale, T.Bruser, A.J.Driessen (2008) Sec- and Tat-mediated protein secretion across the bacterial cytoplasmic membrane--distinct translocases and mechanisms. Biochim Biophys Acta, 1778:1735-1756.
4. C.Robinson, A.Bolhuis (2004) Tat-dependent protein targeting in prokaryotes and chloroplasts. Biochim Biophys Acta, 1694:135-147.
5. N.Gupta, S.Tanner, N.Jaitly, J.N.Adkins, M.Lipton, et al. (2007) Whole proteome analysis of post-translational modifications: applications of mass-spectrometry for proteogenomic annotation. Genome Res, 17:1362-1377.

Prediction of human cilia-related genes by analysis of open-access transcriptomic and proteomic resources

Alexander Ivliev, Marina Sergeeva

Group of Lipid Systems Biology, A. N. Belozersky Institute of Physico-Chemical Biology, Moscow State University, Moscow, 119992, Russia, ivliev.alexa@gmail.com

Cilia are microtubule-rich organelles which protrude from cell surface and play important roles in motility, sensory perception and development in a wide range of eukaryotes including human. Mutations affecting cilia-related proteins are associated with a range of human diseases, known as ciliopathies. To strengthen the basis for understanding molecular biology of the cilia, it is important to identify cilia-related genes at a genome-scale.

We hypothesized that an effective approach to identification of such genes would be analysis of gene coexpression networks (WGCNA algorithm) in human tissues that contain ciliated epithelium: airways, fallopian tubes and brain. 12 large microarray datasets were found for these tissues in the NCBI GEO database (a total of 1,645 tissue samples). In each dataset, we constructed a gene coexpression network using a previously described procedure [1]. This allowed clustering of coordinately expressed genes into modules with specific biological functions (25 modules on average per dataset). In 10 of the 12 datasets, these modules included a reproducible cilia-related module, as revealed by its enrichment with known proteins from the cilium organelle (DAVID, $P < 0.001$). We summarized gene composition of the ciliary module across the datasets, which produced a consensus list of 371 genes (FDR < 0.5%). According to CilDB database, at least 140 of them encode known evolutionarily conserved components of cilia and flagella in eukaryotes. Meanwhile, 74 other genes represented novel ciliary candidates which are predicted as cilia-related for the first time.

To validate the identified genes, we used a novel proteomic resource − Human Protein Atlas [2], which contained antibody-based tissue staining images for 218 proteins from our consensus module. 72% of these proteins, including 25 of the novel candidates, were found to be more abundant in ciliated cells compared to other cell types within airways and fallopian tubes. This strongly supports these proteins as related to cilia. Additionally, the Human Protein Atlas provided detail on the proteins' subcellular localization: cilia (79% of the proteins), apical cytoplasm (16%), whole cytoplasm (4%) and nuclei of ciliated cells (1%). These data can serve as a starting point for functional studies of those proteins whose role in ciliated cells is unknown. For instance, a cytoplasmic enzyme RBKS likely contributes to energy supply for cilia beat, while an apical protein TSGA10 could possibly be associated with basal bodies of cilia. The analysis suggests advantage of combined use of expression profiling and Protein Atlas data for large-scale annotation of protein function.

Little is known about differences in ciliated cells between tissues. To search for such differences, we compared the cilia-related module between brain, airways and fallopian tubes. Four genes were identified as putative tissue-specific components of ciliated cells ($P < 0.05$). The most statistically significant gene was SLC47A2 which belonged to the ciliary module exclusively in the brain tissue, where it furthermore occupied a key hub position. In situ hybridization (ISH) images of brain tissue sections (Allen Brain Atlas) confirmed this gene to be highly expressed specifically in ciliated cells in brain. SLC47A2 is a plasma membrane transporter and could have a secretory function in brain by transporting metabolites, drugs and toxins from nervous tissue into the surrounding cerebrospinal fluid because a similar function has been shown for SLC47A2 in kidney where it transports positively charged substances from blood into urine [3]. Because SLC47A2 in brain is expressed exclusively in ciliated cells but its function seems unrelated to ciliary mechanical movement, this supports the current view that specialized functions of ciliated cells in the human body are broader than only those related to mechanical cilia beat.

Taken together, we identify novel proteins related to the cilium organelle. Our data also support utility of coexpression networks and proteomic resources for studying molecular features of individual cell types in multicellular organisms.

This work is partly supported by grant RFBR 10-04-01385-a.

1. Ivliev AE et al. (2010), *Cancer Res*, **70:**10060-70.
2. Uhlen M et al. (2011), *Nat. Biotechnol*, **28:**1248–50.
3. Masuda S et al. (2006), *J Am Soc Nephrol*, **17:**2127-35.

Secondary Structure Prediction and Molecular Modeling of Human MKP1/ DUSP1

Kaiser Jamil and Sabeena M.

School of Life Sciences, Centre for Biotechnology and Bioinformatics, Jawaharlal Institute of Advanced Studies, Budha Bhawan 6th floor, Secunderabad- 500003, A.P. India, Kaiser.jamil@gmail.com

MKP-1, also known as dual-specificity phosphatase-1 (DUSP1), is a member of a family of proteins that dephosphorylates both threonine and tyrosine residues and thereby serves as a key negative regulator of the MAPK cascade4, a major signaling pathway, involved in neuronal plasticity, autoimmune diseases and cancers being a dynamic immune regulator. The major aim of this study includes functional site prediction, Secondary structure analysis and Comparative /homology studies of MKP-1. We retrieved the human MKP-1 sequence from UNIPROTKB (ID P28562) and analyzed the functional residues and pattern using Swisspdb viewer and ScanProsite. We could identify three different regions in MAPK-1 which are involved in the phosphorylation of protein kinase C, Casein II, and Tyrosine-specific protein. ConSurfDB in ConSurf Server was used for the identification of the specific regions in our target sequence, which predicted the closest protein structure with PDB ID 3EZZ chain A with an E-value of 4e-71. JPRED in ExPasy server was used for the secondary structure prediction, which identified the closest functional protein structure with PDB ID 3EZZ with an E-value of 3e-67. Based on the results obtained from domain analysis and secondary structure prediction, the protein 3EZZ was found to be act as a suitable template for further steps. The three-dimensional structure for MAPK-1 was determined by comparative protein modeling method using the program modeler 9v8. The resultant output model files included the detailed information about the target protein and the model building process, functional annotation, a detailed template selection log, target-template alignment, summary of the model building and model quality assessment. The developed model was used in the optimized mode to minimize energy. The model was validated using SAVES which includes procheck, gave a value of 82.5% by analyzing the stereochemical quality of a the predicted model protein structure by analyzing residue-by-residue geometry and overall structure geometry. Verified 3d results which gave a value of 88.19% by calculating the compatibility of the atomic model (3D) with its own amino acid sequence. GROMACS performed the molecular dynamics and optimization of our predicted model. Finally this study revealed the various residues, patterns and the possible secondary structure present in MKP-1. Since 3d structures for these proteins were not available in PDB, this homology modeling study reports for the first time the 3d structure for MKP-1. Our results further suggest that the modeled structure provides good foundation for functional analysis of experimentally derived crystal structures for MKP-1.

Discovering novel drug-target interactions via superimposition of 3D structures

Olga Kalinina[1], Oliver Wichmann[2], Gordana Apic[2], Robert Russell[2]

[1]*University of Heidelberg, Germany, olga.kalinina@bioquant.uni-heidelberg.de*
[2]*University of Heidelberg, Germany*

Existing interactions between molecules can be used to predict new interactions. In the presented study, we use this principle to suggest new protein-chemical interactions via the network derived from three-dimensional structures. For pairs of proteins sharing a common ligand, we use protein and chemical superimpositions combined with fast structural compatibility screens to predict whether additional compounds bound by one protein would bind the other. The method reproduces 84% of complexes in a benchmark, and we make many predictions that would not be possible using conventional modeling techniques. Within 19,578 novel predicted interactions are 7,793 involving 718 drugs, including filaminast, coumarin, alitretonin and erlotinib. The growth rate of confident predictions is twice that of experimental complexes meaning that a complete structural drug-protein repertoire will be available at least ten years earlier than by X-ray and NMR techniques alone.

Identification of Plant Homologues of Dual Specificity Yak1-Related Kinase 1A

P.A. Karpov, A.V. RaYevsky, S.V. IsaYenkov, S.I. Spivak, Ya.B. Blume

Institute of Food Biotechnology and Genomics, Natl. Acad. Sci. of Ukraine, Osipovskogo str., 2a, 04123, Kyiv-123, Ukraine,karpov.p.a@gmail.com

The animal dual-specificity tyrosine-regulated kinases (DYRKs) comprise a family within the CMGC group [PMID: 15068245**]**, homological yeast Yak1 [PMID: 2558053], *Drosophila* minibrain kinases [PMID: 19844572] and participate in several signalling pathways involved in development and cell homeostasis [PMID: 21048044]. The Dyrk1a plays important role in cell proliferation and brain development. The gene mutations are strongly associated with Down syndrome [PMID: 15068245**]**. Dyrk1a are represented in protista, slime molds, fungi and animals, but the existence of their plant homologues are still unclear.

UniProt (www.expasy.org) query, revealed presence of 18 «Reviewed» sequences of animal DYRKs from *Homo sapiens* (Dyrk1a, Dyrk1b, Dyrk2, Dyrk3, Dyrk4), *Macaca fascicularis* (Dyrk3), *Mus musculus* (Dyrk1a, Dyrk1b, Dyrk2, Dyrk3, Dyrk4), *Rattus norvegicus* (Dyrk1a, Dyrk3), *Xenopus laevis* (Dyrk1a), *X. tropicalis* (Dyrk1a), *Gallus gallus* (Dyrk2), *Drosophila melanogaster* (Dyrk2 (smi35A), Dyrk3), and 2 (Dyrk1, Dyrk2) from *Dictyostelium discoideum* (*Mycetozoa*).

In order to find potential plant Dyrk1a homologues, we have implemented SIB BLAST-search within *Viridiplantae* group. We have used human Dyrk1a (Q13627) catalytic domain as reference sequence. The primary group of plant homologues was 90 potential protein kinases having entire catalytic domain (assigned based on analysis in SMART - http://smart.embl-heidelberg.de/). In comparison with human Dyrk1a, the group exhibits 31-60% sequence identity. At the same time, the similarity was 50-79% and 15% of gaps. All plant sequences are represented by proteins with unknown function, and have «Unreviewed» UniProt status. The Neighbor-Joining clustering of plant homologues and «Reviewed» DYRKs catalytic domains resulted in common clade. Based on common clustering the 34 plant homologues were identified, among them 18 potential plant DYRK-like kinases were represented by complete sequences (Fig).

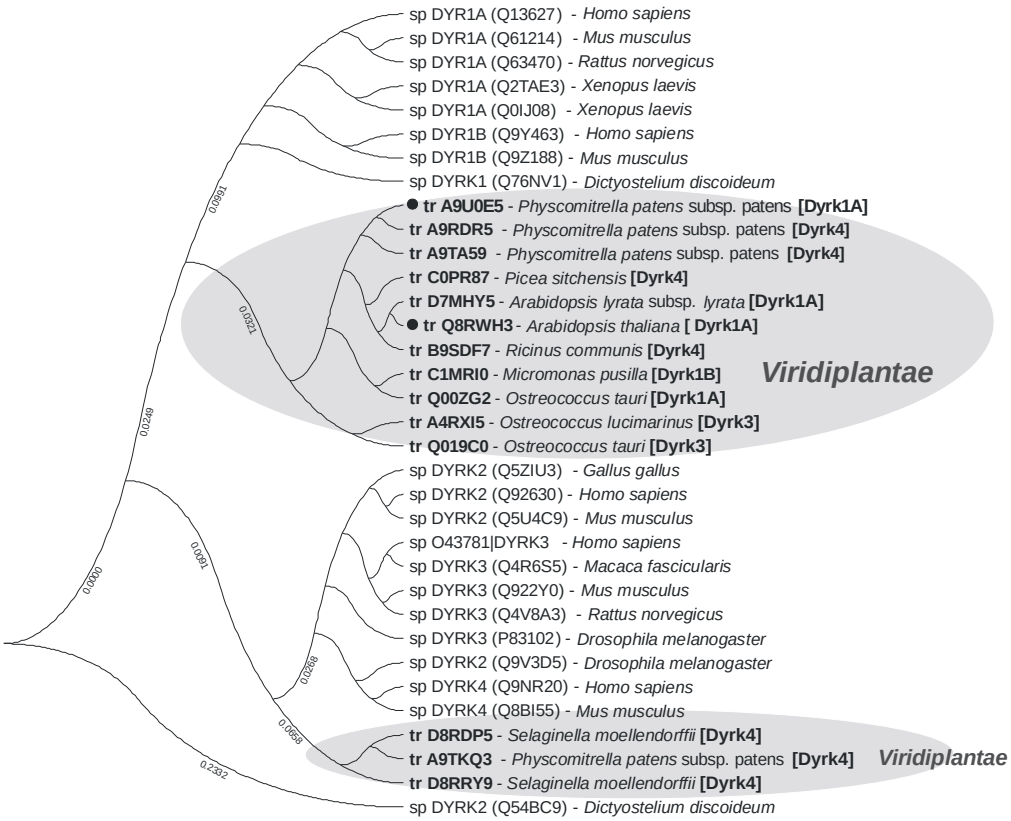


Fig. - Final bootstrap Neighbour Joining analysis *of Dual specificity Yak1-related kinases (DYRKs) and their plant homologs (Viridiplantae): sp - UniProt «Reviewed» sequences; tr - UniProt «Unreviewed» sequences; [DyrkXX] - results of STRING v.8.3* analysis; ● - plant Dirk1A homologs *selected for spatial structure function prediction*

The analysis by complete sequence STRING v.8.3 tool scanning of plant homologues in human proteome confirm their strong relationship to DYRK group. Therefore, Dyrk1A - A9U0E5 from *Physcomitrella patens* subsp. *patens*, D7MHY5 from *Arabidopsis lyrata* subsp. *lyrata*, Q8RWH3 from *A. thaliana*, Q00ZG2 from *Ostreococcus tauri*; Dyrk1B - C1MRI0 from *Micromonas pusilla*; Dyrk3 - A4RXI5 from *O. lucimarinus*, Q019C0 from *O. tauri* и Dyrk4 - A9RDR5 and A9TA59 from *P. patens* subsp. *patens*, C0PR87 from *Picea sitchensis*, B9SDF7 from *Ricinus communis*, D8RDP5 from *Selaginella moellendorffii*, A9TKQ3 from *P. patens* subsp. *patens*, D8RRY9 from *S. moellendorffii*). Dyrk1A homologs from *P. patens* subsp. *patens* (A9U0E5) and *A. thaliana* (Q8RWH3) were selected for further spatial structure and function prediction (see MCCMB'11 Abstracts: Rayevsky et al., 2011).

Using structure-based drug design to promote the development of persistent chlamydial infection treatment.

A. GRISHIN[1,2], M. KRIVOZUBOV[1,2], D. KIRSANOV[1,2,3], S. DANILENKO[2], E. ZAYAKIN[1], D. DAVYDOVA[1], P. VLASOV[4], N. ZIGANGIROVA[1], A. KARYAGINA[1,3]

[1]N.F. Gamaleya Institute for epidemiology and microbiology, Moscow, Russia; [2]Faculty of Bioengineering and Bioinformatics, Moscow State University, Moscow, Russia; [3]A.N. Belozersky Institute of physical and chemical biology, Moscow State University, Moscow, Russia;[4] Center for Genomic Regulation, Barcelona, Spain;
grishin-a1@yandex.ru, akaryagina@gmail.com

Many species of *Chlamydia* genus are intracellular parasites of higher eukaryotes. In humans, chlamydial infection can cause severe respiratory and urogenital diseases, both acute and chronic, the latter one proving to be the most challenging for clinical practice. While commonly used antibiotics successfully suppress initial acute infection, this type of treatment appears to be ineffective against the persistent form of the parasite. Thus, the perspective approach is to target those chlamydial proteins that are responsible for the persistent infection using a well established in the past two decades methodology of virtual ligand screening.

Several proteins that potentially play role in the development of the persistent form of the infection have been described in the literature [1]. Of them, crystal structures are only available for CPAF protease and a series of close homologues of type III secretion system proteins from several related gram-negative bacteria species. Since no known inhibitors for these proteins except covalently binding to CPAF omuralide exist, simple ligand-based screening does not appear to be possible. The structure-based virtual screening was performed against CPAF crystal structure [2], and against homology model of chlamydial CdsN ATPase involved in the effector proteins export by the type 3 secretion system [3]. CPAF protease is highly conserved among various *Chlamydia* species, but lacks significantly close homologues in humans, which suggests the possibility of developing a highly specific inhibitor. On the other hand, the specificity may be a problem for the design of CdsN ATPase inhibitor.

The screening was performed over four libraries of chemical compounds from different vendors. Since the cumulative size of the screened libraries exceeded 1.5 M compounds, the clustering and selecting for only diverse compounds was performed prior to the screening. Selected top scoring ligands and their structural homologues were further subjected to the experimental testing. The test systems include *in vitro* model of chlamydial infection in eukaryotic cells, another system that is capable of discriminating the compounds that specifically inhibit type III secretion system, and yeast cells engineered to express CPAF protease, as well as CPAF substrate cleavage assay.

1. A.S. Karyagina et al. (2009) Effector proteins of Chlamydia, *Molecular Biology (Moscow)*, **43(6)**:963–983.
2. Z. Huang et al. (2008) Structural basis for activation and inhibition of the secreted *Chlamydia* protease CPAF, *Cell Host & Microbe*, **4:**529–542.
3. K. Imada et al. (2007) Structural similarity between the flagellar type III ATPase FliI and F1-ATPase subunits, *PNAS*, **104**:485-490.

Functional annotation of regulons controlled by RNA regulatory elements in complete bacterial genomes

Marat D. KAZANOV[1] , Semen A. LEYN[1,2], Pavel S. NOVICHKOV [3],
Dmitry A. RODIONOV[1,2]

[1]Institute for Information Transmission Problems RAS (Kharkevich Institute), Moscow, Russia; [2]Sanford-Burnham Medical Research Institute, La Jolla, California, USA; [3] Lawrence Berkeley National Laboratory, Berkeley, California, USA
mkazanov@burnham.org, sleyn@burnham.org, psnovichkov@lbl.gov, rodionov@burnham.org

Various regulatory RNA structures including *cis*-acting metabolite-sensing riboswitches, T-boxes, and attenuators and *trans*-acting small RNAs control gene expression without involvement of specific regulatory proteins, such as transcription factors. The mechanism of control of gene expression by *cis*-regulatory RNAs involves formation of alternative mRNA structures that either terminate transcription or inhibit initiation of translation. Different classes of RNA elements use different mechanisms to sense the concentration of a metabolite. Typically, an effector-responsive protein factor specifically binds the *cis*-regulatory RNA that is rather small and simple in structure (e.g., the tryptophan-responsive TRAP protein in *B. subtilis*). A unique class of RNA elements, T-boxes in Gram-positive bacteria, interacts directly with specific uncharged tRNAs to promote expression of target genes in response to amino acid concentrations. Riboswitches are widespread genetic elements with a complex structure that directly sense metabolites and control gene expression of related metabolic pathways. Each riboswitch class is defined by a core of conserved base-paired elements and consensus nucleotides at specific positions and is highly specific to its cognate effector metabolite.

A high level of conservation of primary and secondary structures of riboswitches is very useful for their identification by comparative genome analysis (reviewed in [1]). RNA motifs for 13 known classes of riboswitches, T-boxes, and many experimentally uncharacterized RNA regulatory elements identified in prokaryotic genomes are available within the Rfam database [2]. However, consistent description of regulons and gene functions controlled by these RNA motifs in bacterial genomes is currently not implemented in any public web-resource.

In this work, we mapped 39 known regulatory RNA motifs from the Rfam database and analyzed the genomic context of the respective regulons in 108 microbial genomes from 10 taxonomic lineages. Overall we found and functionally described ~2000 regulatory RNAs that form ~200 regulogs (a set of regulons controlled by individual RNA motif in a single lineage).

The obtained collection of RNA-mediated regulons is available for view in the RegPrecise database of manually curated regulons in prokaryotic genomesc (http://regprecise.lbl.gov) [3].

Among various metabolites recognized by known classes of riboswitches are vitamins such as cobalamin (coenzyme B_{12}), thiamin pyrophosphate (TPP), flavin mononucleotide (FMN); amino acids including lysine, glycine, S-adenosylmethionine (SAM), and S-adenosylhomocysteine (SAH); nucleotides including purines, queuosine, and cyclic di-GMP (binds to GEMM riboswitch). Comparative analysis of the regulon content using the RegPredict web-server (http://regpredict.lbl.gov) allowed us to describe the functional roles of genes controlled by each type of RNA motifs in bacterial genomes. Over a hundred of novel families of genes controlled by the RNA motif regulons have been predicted and their functions have been analyzed in the context of the respective metabolic pathways. For previously uncharacterized RNA motifs we attempted to predict their effector molecule that can induce the respective riboswitches. For instance, we predict that the yybP-ykoY riboswitch is likely modulated by calcium ions and controls the calcium homeostasis in the cell, whereas the ykkC riboswitch was found to control the urea metabolism.

The largest total number of RNA motifs was identified for the T-box motif (450 motifs controlling 19 amino acid-specific regulons in 4 lineages), and for several metabolite-sensing riboswitches including the TPP, cobalamin, SAM, glycine, FMN, *yybP-ykoY*, and lysine riboswitches (90-240 motifs per regulon in 6-10 lineages). The most widespread in the analyzed 10 lineages RNA motifs are the TPP and *yybP-ykoY* riboswitches (found in all lineages), the FMN (9 lineages), the cobalamin and glycine (8 lineages), the GEMM (7 lineages), the SAM, lysine and purine riboswitches (6 lineages). The lineage-specific RNA motifs are the ylbH in Bacillales, the queuosine in Streptococcus, the SAH and sucA in Ralstonia, and the Mg-sensor in Enterobacteriales. Amino acid-specific attenuators (His, Leu, Thr, and Trp) and the S15 leader, controlling the ribosomal gene *rpsO,* are highly conserved in two lineages of gamma-proteobacteria (Enterobacteriales and Shewanella); in addition the His- and Trp-specific attenuators are present in some Thermotogales. The L10, L13, L19, L20, and L21 leaders controlling the ribosomal genes are highly conserved in Firmicutes (Bacillales, Staphylococcus, Streptococcus); the L10 and L20 leaders were also found in Cyanobacteria and Thermotogales.

1. A.G. Vitreschak et al. (2004) Riboswitches: the oldest mechanism for the regulation of gene expression? *Trends Genet*. 20:44-50.
2. P.P. Gardner et al. (2011) Rfam: Wikipedia, clans and the "decimal" release. *Nucl. Acids Res*. 39:D141-5.
3. P. Novichkov et al. (2010) RegPrecise: a database of curated genomic inferences of transcriptional regulatory interactions in prokaryotes. *Nucl. Acids Res*., 38: D111-8.

Evolution study and classification of carbohydrate metabolism genome loci in bacteria

Pavel Shelyakin, Anna Kaznadzey
[1]*IITP RAS, Russian Federation, vzmisha4@gmail.com*

The aim of this study is to explore genome loci of the carbohydrate metabolism in bacteria. Such loci consist of genes encoding proteins which participate in the biochemical transformations of carbohydrates, such as phosphorylation, hydrolysis, isomerisation, etc., and also in the transport and regulation of transcription . Co-localisation of proteins belonging to different, isofunctional families and sub-families allows us to obtain information about evolutionary compatible combinations, and to assess functional compatibility for various proteins.

1. Problem statement

The bacterial carbohydrate metabolism is extremely diverse due to the ability of bacteria to assimilate a wide range of substrates. Genes encoding enzymes adjacent in a methabolic pathway are often located near each other on the chromosome, forming functional loci. Genes from such loci may belong to different families, defined by functional and structural features of the encoded proteins. The aim of this study is to explore the compatibility of groups of proteins that may perform the same function, but belong to different families. Analysis of all possible combinations occurring in bacterial genomes allows us to determine the "standard" compositions of the loci and to classify them. Such classification may be specific for taxonomic groups of bacteria. If members of one family are found in several diverse loci, one may suspect horizontal transfer that should be then confirmed by the phylogenetic-tree analysis. Our aim is also to determine the scope and the specifics of such events.

2. Results and Discussion

The first step was to select an appropriate classification of proteins from the bacterial carbohydrate metabolism. We considered combinations of different classification schemes, including annotated EC-numbers, Pfam-families, COG-families, clustering by bidirectional best hits, and also attempted to divide the obtained families using multiple alignment of their amino-acid sequences within the major functional classes. The most extensive and yet accurate classification system was obtained on using COG families, and about 270 carbohydrate metabolism-related COG families were analyzed further. In large families, subfamilies were identified, based on multiple alignment using Clustalw. Large, well-studied families such as hydrolases and lyases were further structured by the CAZY classification.

Hence, each protein was assigned to a family or subfamily.

Loci were formally defined based on gene proximity. We constructed a graph, where families and subfamilies formed vertices, and arcs connected vertices if members of the respective (sub)families were observed in one locus at least once. We constructed tables with lines and rows corresponding to two sets of the isofunctional (sub)families, and identified highly preferred or avoided pairs of (sub)families. We then weighted arcs by the excess of the number of occurrences, and removed arcs in order of increasing weight. We analyzed the size of the largest connected component dependent on the number of removed arcs, and terminated the process when the giant component dissolved. The obtained components correspond to standard loci, whereas the low-weighting connections between them may stem from horizontal transfer or gene shuffling within a genome followed by change of specificity.

This is joint work with M.Gelfand.

3. Literature

[1] Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M., and Hirakawa, M. KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res. (2010), 38, D355-D360.
[2] Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K.F., Itoh, M., Kawashima, S., Katayama, T., Araki, M., and Hirakawa, M., From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res. (2006), 34, D354-357.
[3] Kanehisa, M. and Goto, S., KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. (2000), 28, 27-30.
[4] Markowitz VM, Chen, I.A., Palaniappan K, et al. The Integrated Microbial Genomes system: an expanding comparative analysis resource. Nucleic Acids Res. (2010) , 38
[5] R.D. Finn, J. Mistry, J. Tate, P. Coggill, A. Heger, J.E Pollington, O.L. Gavin, P. Gunesekaran, G. Ceric, K.Forslund, L. Holm, E.L. Sonnhammer, S.R. Eddy, A. Bateman, The Pfam protein families database. Nucleic Acids Research (2010), Database Issue 38:D211-222
[6] Marchler-Bauer A, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, Fong JH, Geer LY, Geer RC, Gonzales NR, Gwadz M, He S, Hurwitz DI, Jackson JD, Ke Z, Lanczycki CJ, Liebert CA, Liu C, Lu F, Lu S, Marchler GH, Mullokandov M, Song JS, Tasneem A, Thanki N, Yamashita RA, Zhang D, Zhang N, Bryant SH., CDD: specific functional annotation with the Conserved Domain Database. Nucleic Acids Res. (2009), 37, D205-10.
[7] PHYLIP (http://evolution.genetics.washington.edu/phylip.html)
[8] Larkin M.A., Blackshields G., Brown N.P., Chenna R., McGettigan P.A., McWilliam H., Valentin F., Wallace I.M., Wilm A., Lopez R., Thompson J.D., Gibson T.J. And Higgins D.G., ClustalW and ClustalX version 2. Bioinformatics (2007), 23(21): 2947-2948.
[9] Caspi, R., Altman, T., Dale, J.M., Dreher, K., Fulcher, C.A., Gilham, F., Kaipa, P., Karthikeyan, A.S., Kothari, A., Krummenacker, M., Latendresse, M., Mueller, L.A., Paley, S., Popescu, L., Pujar, A., Shearer, A., Zhang, P. and Karp, P.D., The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases Nucleic Acids Res. (2010), 38(1):D473-D479.
[10] Selkov E Jr, Grechkin Y, Mikhailova N, Selkov E. MPW: the Metabolic Pathways Database. Nucleic Acids Res. (1998), 26(1):43-5.
[11] Tatusov RL, Galperin MY, Natale DA, Koonin EV., The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res. (2000), 28(1): 33-36.
[12] Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV., The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res. (2001), 29(1): 22-28.

Modeling of pathway plasticity in cancer.

Alexander Kel

[1]*geneXplain GmbH, Wolfenbuettel, Germany*
[2]*Institute of Systems Biology Ltd., Novosibirsk, Russia*
[3]*Institute of Chemical Biology and Fundamental Medicine, Novosibirsk, Russia*

Massive changes of expression of hundreds of genes as well as changes in genomic and epigenomic landscapes observed in cancer often represent just an "echo" of relatively few causative molecular processes in the cells taking place during the malignant transformation. Non-reversible structural changes in gene regulatory networks may cause carcinogenic transformation of the cell homeostasis switching it from the normal state to the disease state. We call such structural network changes as "pathway plasticity". Analysis of this phenomenon helps us to understand the mechanisms of molecular switches (e.g. between programs of cell death and programs of cell survival) and to identify perspective biomarkers and drug targets of cancer. In the current study we applied systems biology approaches to understand mechanisms of non-genotoxic carcinogenicity. EGF transgenic mouse model, as a model of sporadic liver cancer, was subjected to advanced methods of systems biology based on a combination of sequence analysis and entrained graph-topological algorithms. Promoter analysis of differentially expressed genes in the pre-tumor and the tumor states suggested the majority of regulated transcription factors to display similar activation pattern between these states. Some TFs those exhibit clear specificity to either the pre-tumor or the tumor state. Subsequent search for signal transduction key nodes upstream of the identified transcription factors and their targets suggested the insulin-like growth factor pathway to render the tumor cells independent of EGF receptor activity. Together, we propose an approach for modeling of plastic structural changes in the intracellular regulatory network helping to understand the mechanism of the switch in autocrine signaling which through epigenetic changes in the regulatory regions of target genes leads to non-reversible rewiring in the signaling network to foster tumor growth that was initially triggered by EGF. Deciphering the intimate mechanisms of the malignant transformation using systems biology approaches helping to combat cancer and other diseases is the ultimate goal of the systems medicine.

GeneXplain platform for Systems Medicine.

Tagir VALEEV[1], Anna RYABOVA[1], Nikita TOLSTYH[1], Fedor KOLPAKOV[1], Alexander KEL[2]

[1]*Institute of Systems Biology, Detskiy proezd 15, Novosibirsk, Russia, amaembo@gmail.com*
[2]*geneXplain GmbH, Am Exer 10b, Wolfenbüttel, Germany, alexander.kel@genexplain.com*

We have developed an integrated geneXplain platform (platform.genexplain.com) for systems medicine studies. GeneXplain platform is developed on the basis of the BioUML framework and devoted for causal interpretation of data coming from microarray, proteomics, miRNA and ChIP-chip/seq experiments. GeneXplain platform applies an unique upstream analysis approach based on implementation of machine learning and graph topological analysis algorithms in order to identify causality keynodes in the network of gene regulation and signal transduction [1] and combines it with full genome sequence analysis and chemoinformatics methods for drug discovery.

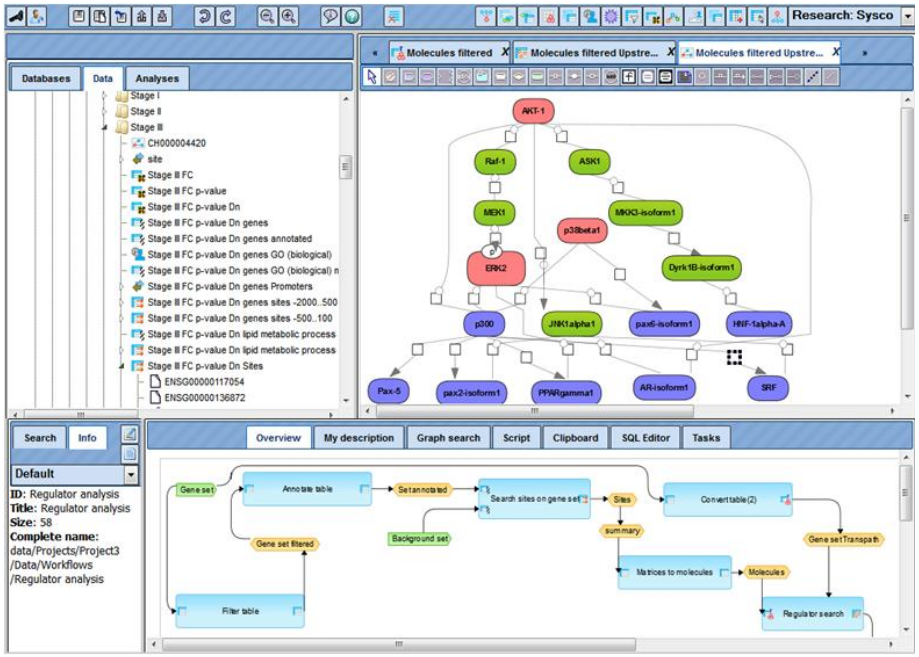


Fig. User interface of the geneXplain platform. The user data and results are in the tree aria. The workflow for identify- cation of master regulators is shown bellow. The reconstructed regulatory diagram shows found keynodes (red nodes) and target transcription factors (blue nodes)

The power of this approach is in identification of causal biomarkers - those which are more than just correlating with disease or treatment outcome, but which are parts of the disease mechanism, which may differ in different patient cohorts. Such personalized networks are analyzed in order to find master regulators - key nodes in the network triggering the disease. Identification of such keynodes allows also to modularize the regulatory networks and build dynamic models of realistic size in order to identify different modes of dynamics of cellular networks discriminating disease from the normal state of the system. Such keynodes and genes directly influenced by them are considered as promising biomarkers which can discriminate

patients into cohorts from the disease mechanism point of view. Another application of such analysis is identification of multiple drug targets in personalized disease networks allowing to guide the choice of the disease treatment by poly-pharmacological drugs.

In a case study on breast cancer, we analyzed a large scale gene expression and ChIP-seq data from a study of cancer samples treated with antineoplastic agents including the novel drug compounds - RITA and Nutlin, targeting p53 and Mdm2. We analyzed promoters of downregulated pro-survival genes and identified combinations of transcription factors involved in their regulation. Topological modeling of the signal transduction network upstream of these transcription factors revealed key-nodes - potent master-regulators of the cell survival program that prevent efficient apoptosis of cancer cells. We considered these key-node proteins (e.g. PI3K subunits) as causal biomarkers as well as prospective targets for novel anticancer drug combinations. We applied a cheminformatics computer tool PASS to these targets and identified two novel prospective antineoplastic chemical compounds which were experimentally validated in a cellular assay confirming their synergistic potential in highly selective triggering of apoptosis of cancer cells.

Acknowledgements. Parts of the work were funded by a grant of the German Ministry of Education and Research (BMBF), GerontoShield and by EU grants Net2Drug, LipidomicsNet and SysCol.

1. Stegmaier P, Voss N, Meier T, Kel A, Wingender E, Borlak J. (2011) Advanced computational biology methods identify molecular switches for malignancy in an EGF mouse model of liver cancer., PLoS One, **6**(3):e17738..

A molecular survey across lifespan: human brain evolution and aging.

Philipp Khaitovich[1,2]

[1]*CAS-MPG Partner Institute for Computational Biology, Chinese Academy of Sciences, 320 Yue Yang Road, 200031 Shanghai, China*
[2] *Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, 04103 Leipzig, Germany,*
khaitovich@eva.mpg.de

Phenotypically, humans stand out from other primate species in many respects. Here, we focus on two characteristics specific to humans: unique cognitive abilities and extended lifespan. To approach these questions, we surveyed RNA, protein and metabolic changes across the human lifespan and compared them to the changes in other species: chimpanzees, macaques and mice.

The results point out to changes in expression of specific pathways, possibly underlying human-specific cognitive abilities. Further, we show that lack or regulation control, rather than accumulation of stochastic damage might be responsible for the detrimental effects of aging.

Evolution of diversity in Ubiquitin conjugating enzymes

Muhummadh Khan[1], Kaiser Jamil[2]

[1]*Mahaveer Hospital and Research Center, Masab Tank, Hyderabad, A.P., India, muhummadh@gmail.com*
2Dean, School of Biotechnology, JNiAS, Buddha Bhavan, Hyderabad

Ubiquitin conjugating enzymes (E2) have been implicated in cancer. Intriguingly these enzymes are also part of important pathways required for normal functioning of a cell. We were interested in their evolution which would shed light on how these enzymes acquired their function. So we generated a phylogenetic tree for these enzymes. In order to reduce redundancy of sequences in the tree, we selected only representative sequences from a group of highly similar homologs. The phylogenetic tree was constructed using the maximum likelihood method.

We found that E2 enzymes were evolving irrespective of species evolution and they were under selective pressure to maintain their structure. The phylogenetic tree of ubiquitin conjugating enzymes showed that these enzymes were clustered into clear subclasses (see figure). Ubiquitin like proteins (SUMO and NEDD) conjugating enzymes were seen as one of the subclasses in the tree. These different subclasses of E2s cater to different modifier activating enzymes (E1). The ubiquitin conjugating enzymes are exclusive to eukaryotes. They are absent in prokaryotes. If we consider another clade of NEDD8 conjugating E2s, it is seen that these are grouped regardless of species evolution. Another fact which stands out from the tree is that the similarity sharing E2 enzyme clades are specifically involved in a given process. These cellular processes are diverse ranging from cell cycle, transcript regulation, proteolyses to specific proteolyses.

Hence we assume that evolution of the E2 enzymes was destined to specific functions in cell. This is further supported by the fact that different clades of the phylognetic tree show functionally similar E2s from different organism group as one. It indicates that E2s have evolved different lineages destined to cater to different pathways functioning in a cell. We expect that the lineages of E2 enzymes diversified irrespective of the organismal evolution. This is indeed interesting to note that different paths chosen in biomolecular evolution may or may not be in accordance with the evolution of the species. It seems like these two processes happen to their own ends irrespective of the consequence at a different level.

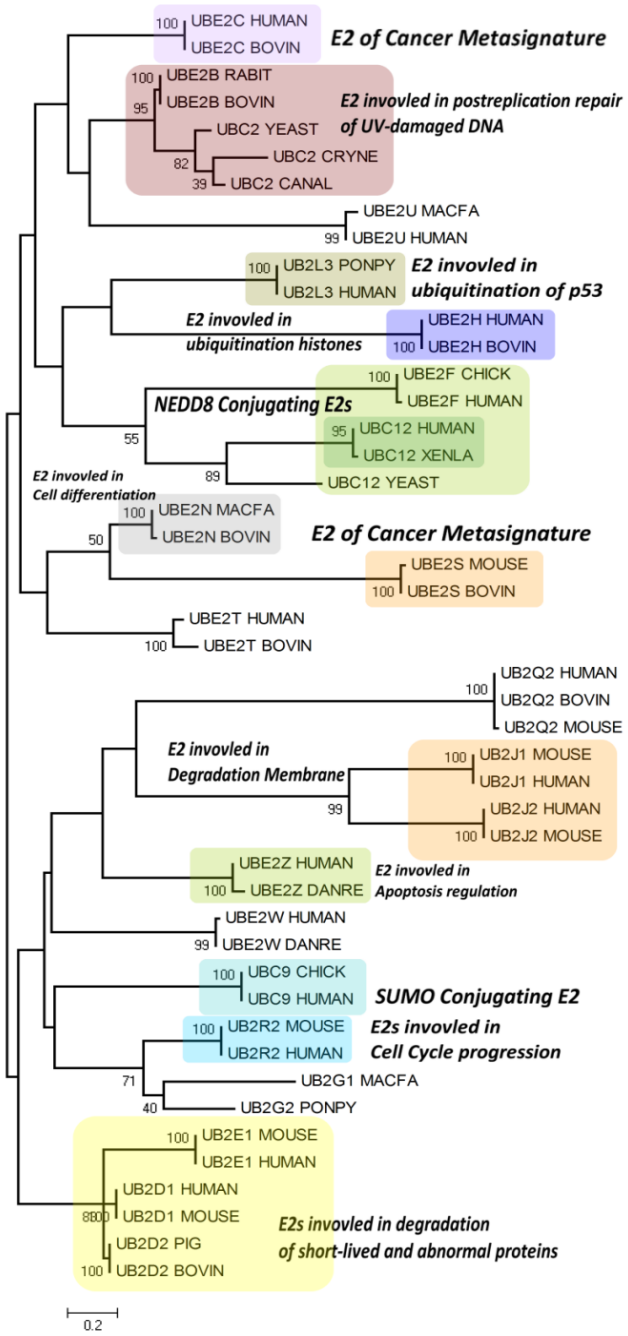


***Figure**: Phlyogentic tree of Ubiquitin conjugating enzymes*

Bioalgorithm development for virulence screening.

Khaled Khanchouch[1], Mohamed Rabeh Hajlaoui[2], Elena Ustymovych[3], Hakan Kutucu[4]

[1]*University of Tunis, Tunisia, Tunisia, khanchouchkhaled@yahoo.fr*
[2]*Institute of Agronomic Research of Tunisia, Tunisia, Tunisia, hajlaoui.rabeh@iresa.agrinet.tn*
[3]*IT Department of Informatization Center, Ukraine, Elena.ustymovych@gmail.com*
[4]*Izmir Institute of Technology, Turkey, Turkey, hakankutucu@iyte.edu.tr*

Statistical analyses for virulence screening to discriminate between pathogenic microbes are based on relative comparing methods. They allow the classification of the tested isolate and situate it inside relative groups of virulence. The same isolate compared to another different samples originated from different geographical area can have less or much higher degree of virulence. As many diseases are known by their cosmopolite repartition, universal tool and a standard method to determinate the virulence level of the pathogen are needed. For this reason a bio-algorithm is developed to evaluate the proper virulence of each tested isolate of the example pathogenic fungal population of Phoma tracheiphila, collected from different regions.

Numerical analyses of the experimental results reveal that the inner virulence of each tested isolate can be simulated via a polynomial model of fifth degree. A bio-algorithm based on this mathematical simulative model is built to find out the specific polynomial functions of the pathogenic studied population. In order to define the representative polynomial's parameters of each tested isolate a linear equations system is solved by Gaussian elimination method (Carl Friedrich Gauss, 1810) [1]. The polynomial's coefficients determined by the bio-algorithm are in concordance with its obtained with Mathematica uses Hermite interpolation technique to find fitting curves to a given sets of data [2]. Since our algorithm is based on Gauss elimination method, the time complexity of the algorithm is O(n3).

[1] Gareth Williams, Linear Algebra with Applications, Seventh Edition, Jones & Bartlett Publishers, 2009, 554 pages.
[2] Stephen Wolfram, The MATHEMATICA ® Book, Version 4, Cambridge University Press, 1999, 1469 pages.

Interpreting chromatin states in model organisms

Peter Kharchenko, Peter Park
Harvard Medical School, United States, peter.kharchenko@post.harvard.edu

The NIH model organism Encyclopedia of DNA Elements (modENCODE) project is focused on using the *D. melanogaster* and *C. elegans* model systems to build a foundation for understanding genome function by providing the community with a comprehensive map of the distributions of chromatin components, transcription factors, transcripts, small RNAs, and origins of replication[1,2]. The packaging of DNA into chromatin is of particular interest, as its conformations impact chromosome function, gene regulation, and other key cellular processes. To systematically explore the functional properties of different chromatin configurations, we have utilized a Hidden Markov Model approach, identifying prevalent combinatorial patterns of histone modifications within the *D. melanogaster*[3] and *C. elegans* genomes that most readily explain the observed chromatin variability. For every examined cell type, the models associate each genomic location with a chromatin particular state, generating a chromatin-centric annotation of the genome. We then examined each state for association with non-histone proteins, chromatin accessibility, transcriptional features and known functional elements.

We find that in both organisms, chromatin state composition distinguishes multiple categories of both transcriptionally active and silent genes. Some of the variation can be attributed to presence of large-scale chromatin features demarcating significant portions of the chromosomes. In flies such regions are limited to the pericentromeric and subtelomeric chromatin regions and two exceptional chromosomes (4^{th} and X). By contrast, the chromatin organization of the worm genome contains much broader domains with persistent chromatin differences, separating active genes into three major classes that are linked to chromosome centers, arms and X chromosome. We examine binding of chromatin modifying enzymes associated with the distinguishing chromatin features within these classes.

In addition to the differences linked to genomic context, chromatin composition differentiations a subset of genes with high fraction of intronic sequence. Such genes, commonly associated with regulatory functions, contain one or more regions marked by an unusual combination of histone aceytlations and methylations, binding of specific chromatin remodeling machinery, and very high nucleosome turnover. Such intron-biased regions account for the majority of the DNase I hypersensitive sites outside of the promoter regions, suggesting their regulatory role.

The presence of distinct gene subclasses is not limited to transcriptionally active genes. Genes within Polycomb domains can also be found in several distinct states. We show that in addition to previously characterized 'balanced' state that allows for productive transcription in the presence of the Polycomb complexes, fly genomes also maintain a number of important regulatory genes in a transcriptionally paused state[3]. The chromatin signature associated with such pausing is distinct from what has been previously characterized in mammals[4], and the presence of pausing appears to be linked to the proximity to the Polycomb response elements.

Acknowledgements: most of the analysis is based on the data generated by the modENCODE fly and worm chromatin groups, including laboratories of Gary Karpen, Sarah Elgin, Mitzi Kuroda, Vincent Pirrotta, Jason Lieb, Julie Ahringer and Susan Strome; chromatin accessibility data were generated in collaboration with John Stamatoyannopoulos group; transcriptome data generation was lead by Susan Celniker and Robert Waterston.

References:

1 Gerstein, M. B. *et al.* Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. *Science* **330**, 1775-1787.
2 modEncode *et al.* Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science* **330**, 1787-1797.
3 Kharchenko, P. V. *et al.* Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. *Nature* (2011).
4 Bernstein, B. E. *et al.* A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125**, 315-326 (2006).

Individual differences in gene expression in liver and kidney (Sus scrofa)

Nataliya Khlopova, Tatiana Glazko

Russian State Agrarian University – MTAA, 127550, Timiryasev st., 49, Moscow, Russia,
hns86@mail.ru, vglazko@yahoo.com

DNA microarray technology allows the measure of the mRNA expression level of thousands of genes simultaneously in a single assay. But analysis of gene expression profiling involves such difficulties as effect of cross-hybridization, statistical and technical errors. Another problem is natural variability of biological systems, related with action of endo/exogenous regulatory factors that can increase or decrease gene expression. All this complex of errors reduces to imperfect outcome hybridization reproducibility and hinder from obtaining accurate results. The present work presents comparative analysis of gene expression profiles (GEP) in

liver and kidney of pigs, analysis of genes with different expression level between investigated animals, comparisons of results obtained in our research with the results published by Zhao [1] and analysis of possible reasons of imperfect outcome reproducibility in DNA microarray experiments carried out by different research groups.

The mRNA expression profiles of liver and kidney of 5 landrace gilts were determined using Swine Protein Annotated Microarrays targeting 15204 genes. Analysis of normalized expression levels were carried out with Exel macro and STATISTICA 7 software for statistical analysis. Information regarding to the gene functions were based on the NCBI database. Metabolic pathways of the genes were determined with KEGG pathway database.

Analysis of the genes brightly different in hybridization intensity (>20000 fluorescent units) between liver and kidney allowed forming group of 24 genes with variable expression level among investigated animals. These genes were subdivided into functional groups with inner statistically significant correlations between expression levels for different animals. As a rule, all genes inside of these functional groups belonged to the general metabolic way. Totally 8 groups were allocated - 6 genes participating in formation of blood clot; 6 genes - in transport and lipid metabolism; 5 genes - markers of blood cells and lysosomes; 3 genes - participants in apoptosis process, 3 glucocorticoid hormone depended genes. Three groups involved two genes each: products, participating in Ca^{2+} transport, in intercellular matrix creation, reflecting the mitochondrion functioning. Comparative analysis of liver GEP obtained in our research with data published by Zhao confirmed the problem of imperfect outcome reproducibility in microarray-based experiments. Among 135 genes, presented as organ-specific in Zhao research, hybridization level of 26 genes didn't exceed the average back-ground signal in our experiment, accordingly, these genes can not considered as organ-specific for liver. According to the hybridization intensity of the rest 51 genes group of 26 genes with constitutive expression and group of 25 genes with variable expression level among investigated animals were formed. These 25 genes parted into 8 functional groups with inner statistically significant correlations between expression levels for different animals: 5 genes participating in formation of blood clot, 8 genes – transporters, 6 substrate dependent genes, 3 genes connected with mitochondria functioning, 3 genes participating in cell division and apoptosis, 3 immune system genes, 3 genes - precursors of Krebs cycle and 2 genes involved in alcohol detoxication. In addition the KEGG pathway analysis revealed on average 1.5 and 3 pathways per gene in the group of genes with and without variable expression level among investigated animals accordingly. These results show the tendency that genes with variable expression level participate in more pathways then genes with constitutive expression. Analysis of 25 genes, presented on our microarray with more

than one probe, allowed forming group of 3 genes with constitutive hybridization signal for all probes, but gene MEP in our research was referred to kidney GEP and hybridization intensity of some probes of the gene HMGCS1 didn't exceed the average back-ground signal. The rest 22 genes had more than twofold differences between hybridization intensity of different probes of the same gene that can testify the presence of the cross-hybridization effect.

The obtained data testifies the necessity of forming constitutive and variable part of GEP. Revealed individual variability of a gene expression between animals is co-ordinated and basically defined on distinctions between animals in regulation of separate biochemical ways by exo/endogenous factors. Imperfect outcome reproducibility in GEP can be explained by the fact that biochemical bases of forming of each gene function depend on gene interaction, resulting in functional network, and each organ-specific function can be affected by increased expression of different genes of such network.

1. Zhao S.H., et al. (2005) Validation of a first-generation long-oligonucleotide microarray for transcriptional profiling in the pig, *Genomics.*, **86(5):**618–625.

Application of a polarizable force field to calculations of relative protein-ligand binding affinities

Oleg Khoruzhii[1], Mikhail Olevanov[2], Vladimir Ozrin[3], Oleg Butin[4]

Russian Federation, bokonon@yandex.ru

For the first time, an explicitly polarizable force field based exclusively on quantum data is applied to calculations of relative binding affinities of ligands to proteins. Five ligands, differing by replacement of an atom or functional group, in complexes with three serine proteases — trypsin, thrombin and urokinase-type plasminogen activator — with available experimental binding data are used as test systems. A special protocol of thermodynamic integration was developed and used to provide sufficiently low levels of systematic error along with high numerical efficiency and statistical stability. The calculated results are in excellent quantitative (RMSD = 1.0 kcal/mol) and qualitative (R^2 = 0.90) agreement with experimental data. The potential of the methodology to explain the observed differences in the ligand affinities is also demonstrated.

Interlaboratory and interplatform comparisons of 117 mRNA and genome sequencing experiments

Ekaterina Khrameeva[1], Mikhail Gelfand[2]

[1]*Moscow State University, Russian Federation, khrameeva@genebee.msu.ru*
[2]*Institute for Information Transmission Problems, Russian Federation, gelfand@iitp.ru*

High-throughput sequencing of whole genomes and transcriptomes has become a major focus of modern biology as DNA sequencing is now available to many more projects, and even single research groups. As the performance of platforms or versions may differ, it is not clear to what extent the obtained results are consistent across platforms or versions thereof, or even between different laboratories [1]. Here, we compared results of 117 human mRNA and genome high-throughput sequencing experiments performed on Illumina and SOLiD platforms of all generations in 21 institutions all over the world. Starting with the raw data, we calculated coverage profiles that are normally used to determine biologically relevant parameters such as gene expression or exon inclusion rates.

The data were retrieved from the NCBI Sequence Read Archive (SRA, http://www.ncbi.nlm.nih.gov/Traces/srahttp://www.ncbi.nlm.nih.gov/Traces/sra). Experiments that had the total amount of sequenced data exceeding 500 million bases (Mb) and were publicly available as of 10th of October, 2010 were selected for further analysis. For each experiment, per-nucleotide gene coverage was calculated. Only single-exon genes were considered to make gene coverage profiles comparable between mRNA sequencing experiments for different tissues that may produce alternative splice isoforms, and also to allow for the comparison of genomic and transcriptomic data. For each gene, the Pearson correlation coefficient of the coverage profiles was calculated between all possible pairs of sequencing experiments.

Clustering of experiments showed that mRNA sequencing experiments coming from the same laboratory tend to have quite similar gene coverage profiles even if such very different tissues as brain and liver were sequenced. At the same time, sequencing of transcriptomes in different laboratories, even from the same tissue and on identical platforms, yielded quite different gene coverage profiles. The genome sequencing experiments do not produce such a clear picture. While some of them cluster by laboratory, others do not.

The observed high dependency of the gene coverage profiles on the producing laboratory demonstrates that comparisons of sequencing results are difficult. This problem is crucial for mRNA sequencing experiments as the goal of most such studies is to perform the comparison of the expression level in various tissues, diseased and healthy individuals, case and control experiments, etc. Studies involving comparative analysis of sequencing data produced by the same laboratory are largely free of such artifacts, whereas biological implications of wider comparisons may require additional controls and normalization procedures.

The work was funded by Dynasty foundation and RFBR grants 10-04-00783 and 09-04-92745.

1. M.Kircher, J.Kelso (2010) High-throughput DNA sequencing – concepts and limitations, Bioessays,32(6):524–36

Factors Affecting Target Site Selection for Drosophila melanogaster LTR-retrotransposons and Retroviruses

Lidia Nefedova, Felix Urusov, Alexander Kim
Department of Biology, Moscow State University, Russian Federation, aikim57@mail.ru

Mobile genetic elements (MGEs) such as retrotransposons with long terminal repeats (LTR-retrotransposons) show significant structural similarity to retroviruses integrated in a host genome (proviruses). It is supposed that life cycles of retroviruses do not differ from those of LTR-retrotransposons. Hence their integration should most likely occur in the same way. The efficiency of retroviral DNA integration depends primarily on the efficiency of the integrase binding with LTRs and with a target DNA. There is so far no clear understanding of the specificity of integration of retroviruses and MGEs in a target site. A variety of factors are supposed to potentially influence DNA target selection, including the transcriptional status of DNA, methylation, association with histones and other DNA-binding proteins, DNA bending.

Gypsy is one of the most studied LTR-retrotransposons of Drosophila. Based on its genetic and biochemical properties *gypsy* is classified as an endogenous retrovirus [1]. Infectious retrotransposons and similar MGEs were placed in the *gypsy* group and named errantiviruses (endogenous retroviruses of insects) [2]. At the present time 26 LTR-retrotransposons are classified as the *gypsy* group elements, some of them are real retroviruses [3]. Thus, *D. melanogaster* is suitable for genetic studies of retroviruses.

We performed a comprehensive computer analysis of the integration specificity of *D. melanogaster* LTR-retrotransposons and retroviruses. We analyzed the genome environment of LTR-retrotransposons, the nucleotide composition of LTR end sequences (including A-philicity, DNA-bending and protein-induced deformability), and structural features of integrases. We classified LTR-retrotransposons based on integration specificity and a comparative analysis of integrase amino-acid sequences. Our analysis revealed that retroviruses of the *gypsy* group and their derivative LTR-retrotransposons show specificity on target DNA, and this specificity differs for representatives of the three *gypsy* subgroups: *gypsy*, *ZAM* and *Idefix*. Integration specificity correlates with target DNA features such as A-philicity, DNA-bending and protein-induced deformability. At the same time, some LTR-retrotransposons do not show specificity of integration. These retrotransposons are related to the *copia* and *BEL* groups, and the *blastopia* and *412* subgroups of the *gypsy* group. We demonstrate – at least for retrotransposons of the *blastopia* and *412* subgroups – the presence of chromodomain in structures of integrases that can promote process of non-specific integration.

This work was supported by the RFBR (grant No. 11-04-00403) and the Federal Program "Scientific and Pedagogical Specialists of Innovative Russia" for 2009-2013.

1. A.Kim et al. (1994) Retroviruses in invertebrates: the gypsy retrotransposon is apparently an infectious retrovirus of Drosophila melanogaster, *PNAS USA*, **91**:1285–1289.
2. J.D.Boeke et al. (2006) Index of Viruses – Metaviridae, In: *ICTVdB – The Universal Virus Database, version 4*. C.Büchen-Osmond (Ed.), N.Y. USA.
3. L.N.Nefedova, A.I.Kim (2009) Molecular phylogeny and systematics of Drosophila retrotransposons and retroviruses, *Mol. Biol. (Mosk.)*, **43**(5):747–756.

NPIDB, a database of structures of nucleic acid – protein complexes

Dmitry Kirsanov[1], Olga Zanegina[1], Andrei Alexeevski[1], Sergei Spirin[1], Anna Karyagina[2]

[1]*Belozersky Institute of Moscow State University, Russian Federation, ddk@belozersky.msu.ru*
[2]*Gamaleya Institute of Epidemiology and Microbiology, Russian Federation, akaryagina@gmail.com*

The resource NPIDB (Nucleic acids – Protein Interaction DataBase) [1] includes a collection of files in the PDB format containing structural information on DNA-protein and RNA-protein complexes, and a number of online tools for analysis of the complexes. Those tools are: an original program CluD [2] for analysis of hydrophobic clusters on interfaces, program for detecting potential hydrogen bonds and water bridges, visualization of structures with Jmol (http://jmol.sourceforge.net/). Information on SCOP [3] and Pfam [4] domains presented in protein chains of structures is presented.

Structures of protein – nucleic acid complexes are extracted from PDB as files in the PDB format. There are two types of PDB files: first, representing asymmetric units (PDB entries "as is") and second, representing biological units. Some structures are revised to correct possible mistakes, such as duplication of atoms, and inconvenience, such as two or more variants of a structure posed in one coordinate space (see, for example, PDB entry 1QPI, where two variants of each DNA chain are superimposed). All structural files of NPIDB are available for download.

Update of the content is done regularly by a special program module. At April 2011, NPIDB contains 3122 structures.

At 2011, a new design of web-interface to NPIDB is created, it is available via Internet: http://npidb.belozersky.msu.ru . The content of the database is presented in three lists: the list of available structures, the list of Pfam protein domains, and the list of SCOP protein domains.

Each individual complex has its own web page, containing general information, links to other resources (e.g., PDBsum [5] and BIPA [6]), a table describing biological units or (in case of structures solved with NMR) models, tables describing Pfam and SCOP domains in protein chains, and the list of available actions (including Jmol visualization).

The detection of Pfam domains is done using the HMM profiles downloaded from Pfam server. Each family has its own web page with the list of entries that include domains of the family. Best resolution representatives of Pfam families are available for download or viewing with Jmol.

List of SCOP domains occurred in DNA-protein and RNA-protein complexes is organized in tree-like form, according to the SCOP classification. For each SCOP family a web-page is created. It contains the list of entries that include domains of the family.

For SCOP families of DNA-binding proteins, the web page also contains a superimposition of domain structures, and the structure-based alignment of the domains. At the alignment, secondary structure as well as residues that contact with DNA via hydrogen bonds, water bridges and hydrophobic interactions are shown. Such families are classified according to the mode of protein-DNA interaction. Basing on interacting elements of protein (helix, sheet, strand, turn) and DNA (the major/minor groove, sugar-phosphate backbone), 20 types of protein-DNA interaction were suggested.

The work is partly supported by the joint grants of Russian Foundation of Basic Research and German Scientific Society no. 09-04-92743 and 11-04-91340, and the grant of RFBR no.10 07 00685

1. S.Spirin et al. (2007) NPIDB, a Database of Nucleic Acids–Protein Interactions. Bioinformatics 23(23):3247–3248.
2. A.Alexeevski et al. (2003) CluD, a program for determination of hydrophobic clusters in 3D structures of protein and protein-nucleic acid complexes, Biophysics 48 suppl. 1, S146–S156.
3. A.G.Murzin et al. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. 247:536–540.
4. R.D.Finn et al. (2010) The Pfam protein families database, Nucleic Acids Research 38, Database issue, D211-222.
5. Laskowski RA. (2001) PDBsum: summaries and analyses of PDB structures. Nucleic Acids Research 29(1):221-222.
6. Lee S, Blundell TL. (2009) BIPA: a database for protein-nucleic acid interaction in 3D structures. Bioinformatics 25(12):1559-1560.

Metagenomic analysis of exoelectrogenic bacteria that power microbial fuel cells

Larisa Kiseleva, Igor Goryanin
Okinawa Institute of Science and Technology, Japan, larisa.kiseleva@oist.jp

The interest of our study, exoelectrogenic microorganisms, form communities that power microbial fuel cells. Systematic investigation of such community structures, genomes and metabolism will provide new opportunities for the sustainable production of energy from biodegradable compounds.

We will describe the computational part of work involved into metagenome studies: universal primers design, identification of microbial diversity, sequence data assembling, metagenome data visualization.

The mRNA characteristics potentially involved in recognition of non-AUG start codons in yeast mRNAs

Oxana Volkova, Alex Kochetov

Institute of Cytology and Genetics SB RAS; Novosibirsk State University, Russian Federation, ak@bionet.nsc.ru

Initiation of translation of most eukaryotic mRNAs is likely to occur by linear scanning (Kozak, 2005). According to the scanning model, 40S ribosomal subunits initiate translation at the first AUG they encounter. The initiation/scanthrough ratio depends on both the AUG context and the features of downstream mRNA fragment. It was also recently shown that a considerable number of *S. cerevisiae* mRNAs (~150) contained alternative non-AUG start codons (Ingolia et al., 2009). It is likely that alternative nonAUG start codons could make an important contribution to eukaryotic proteomes. However accurate prediction of nonAUG-starts demands detailed investigation of mRNA features and currently it is not possible.

Comparative analysis of yeast nonAUG- and AUG-started ORFs (taken from (Ingolia et al., 2009)) has been conducted. These ORFs were located within the 5'-UTR region of mRNAs (upstream ORFs, uORFs) and their functions (if any) remained unknown. However, active initiation of translation at these sites was experimentally verified. Interestingly, nonAUG-started uORFs were characterized by mRNA features facilitating its recognition (very strong nucleotide context, tendency to have unfolded 5'-UTR segment and stable hairpins at CDS positions 14,15). The features of mRNAs with AUG-started uORFs were found to be inhibitory (poor nucleotide context, no signs of optimization of RNA secondary structure). Also, non-AUG-started uORFs considerably more frequently encode polypeptides of larger size. It is likely that AUG-started uORFs are much more frequently used as a regulatory signals and corresponding peptide has no functional significance whereas functions of nonAUG-started uORFs more frequently associated with their protein products.

This work was supported by Russian Foundation for Basic Research (09-04-92653), Russian Ministry of Science & Education (2.1.1/10551 (2.1.1/6382); 02.740.11.0705; P128), and the Program of Russian Academy of Sciences (Molecular and Cellular Biology). The authors are also grateful to SD RAS Complex Integration Program for partial support.

1. M. Kozak (2005). Regulation of translation via mRNA structure in prokaryotes and eukaryotes, *Gene*, **361**: 13-37.
2. N.T. Ingolia, S. Ghaemmaghami, J.R. Newman, J.S. Weissman (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling, *Science*, **324(5924)**:218-23.

Ligands of Adipose Stem Cell Receptors Isolated by High-Throughput Combinatorial Peptide Library Screening

Mikhail Kolonin[1], Anna Sergeeva[2]

[1]*University of Texas Health Science Center at Houston,, United States, Mikhail.G.Kolonin@uth.tmc.edu*
[2]*M.D. Anderson Cancer Center, United States, asergeev@mdanderson.org*

White adipose (fat) tissue has been recently recognized as a source of multipotent mesenchymal stem cells. These adipose stem cells (ASC) may be useful for regenerative medicine applications (1). However, as we recently showed, they can also promote cancer progression (2). The current lack of surface molecules specifically expressed by ASC limits the capacity to assess the role of these cells in tissue repair and pathology. We have previously used in vivo phage display library screens to identify ligands of markers expressed on specific cell populations and used these ligands to develop agents useful for therapeutic cell targeting and imaging (3, 4). Here, we have combined library screening in mice with the methods we developed for ASC isolation by fluorescence-activated cell sorting (FACS) from organ cell suspensions (5). We isolated a number of peptides that specifically home to ASC. Based on the combination of advanced bioinformatics approaches and biochemistry, we used these peptides to purify the corresponding ASC surface receptors. We identified one of the receptors as a previously unreported cleavage product of decorin lacking the glycanation site. We demonstrate that this new DCN isoform is differentially expressed on ASC cell surface. In a screen for proteins binding this DCN isoform, we identified resistin, an adipokine for which the receptor has been unknown. Identification of other ASC ligand / receptor systems is underway, which will lead to a cell surface protein interaction map in white adipose tissue.

We propose to use ASC-homing peptides for targeted therapy delivery based on our established approach (6). Because ASC represent a critical factor underlying the resistance to weight stabilization and the primary driver of weight regain following obesity interventions, depletion of the ASC population with peptide-delivered therapies represents a new approach to long-term control of adipose tissue mass. Moreover, ASC targeting through peptides identified on our studies may become useful for cancer diagnosis and prognosis as well as a complementary cancer therapy.

1. M.G. Kolonin and P.J. Simmons (2009) Combinatorial stem cell mobilization, Nature Biotechnology, 27: 252-253.
2. Y. Zhang et al. and M.G. Kolonin (2009) Mouse white adipose tissue is a source of cells that are recruited by tumors and promote cancer progression, Cancer Research, 15: 5259-5266.
3. W. Arap, M.G. Kolonin et al. (2002) Steps toward mapping the human vasculature by phage display, Nature Medicine, 8: 121-127.
4. A. Sergeeva, M.G. Kolonin et al. (2006) Display technologies: application for the discovery of drug and gene delivery agents, Advanced Drug Delivery Reviews, 58: 1622-1654.
5. Y. Zhang et al. and M.G. Kolonin (2011) Influence of BMI on Level of Circulating Progenitor Cells, Obesity, online, ahead of print.
6. M.G. Kolonin et al (2004) Reversal of obesity by targeted ablation of adipose tissue, Nature Medicine, 10: 625-632.

BioUML – open source plug-in based platform for bioinformatics: invitation to collaboration

Fedor A. KOLPAKOV[1,2*], Nikita I. TOLSTYKH[1], Tagir F. VALEEV[1,3], Ilya N. KISELEV[1,2], Elena O. KUTUMOVA[1,2], Anna RYABOVA[1,3], Ivan S. YEVSHIN[1,4], Alexander E. KEL[1,5]

[1] *Institute of Systems Biology (Biosoft.Ru), Novosibirsk, Russia*
[2] *Design Technological Institute of Digital Techniques SB RAS, Novosibirsk, Russia*
[3] *Institute of Informatics Systems SB RAS, Novosibirsk, Russia*
[4] *Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia*
[5] *geneXplain GmbH, Wolfebuttel, Germany*
fedor@biouml.org

BioUML (http://www.biouml.org) is an open source plug-in based Java platform for bioinformatics, systems biology and biomedical research. It spans the comprehensive range of capabilities including access to databases with experimental data, tools for formalized description of biological systems structure and functioning, as well as tools for their visualization, simulation, parameters fitting and analyses. Due to scripts (R, JavaScript) and workflow support BioUML platform provides powerful possibilities for analyses of high-throughput data. Plug-in based architecture (Eclipse run time from IBM is used) allows to add new functionality using plug-ins. BioUML platform consists of 3 parts:

- BioUML server - provides access to data and analyses methods installed on the server side for BioUML clients (workbench and web edition) via the Internet.
- BioUML workbench - Java application that can work standalone or as "thick" client for BioUML server.
- BioUML web edition - "thin" client for BioUML server (you just need to start web browser) that provides most of functionality of BioUML workbench. It uses AJAX and HTML5 <canvas> technology for visual modeling and interactive data editing.

We hope that BioUML platform will allow to join efforts of many bioinformaticians in Russia and worldwide. BioUML platform allows bioinformaticians to concentrate on solving of their scientific problems and frees them from many routine tasks related with data access, conversion of formats, security, access rights, etc. We invite third party developers to integrate they algorithms and software into BioUML platform. For this purpose we are providing BioUML Development Kit (BDK) that allows to Java developers implement their own plug-ins for BioUML workbench and BioUML server, initial version of BioUML developer guide and a number of examples. BioUML team will be happy to help you and answer on your questions.

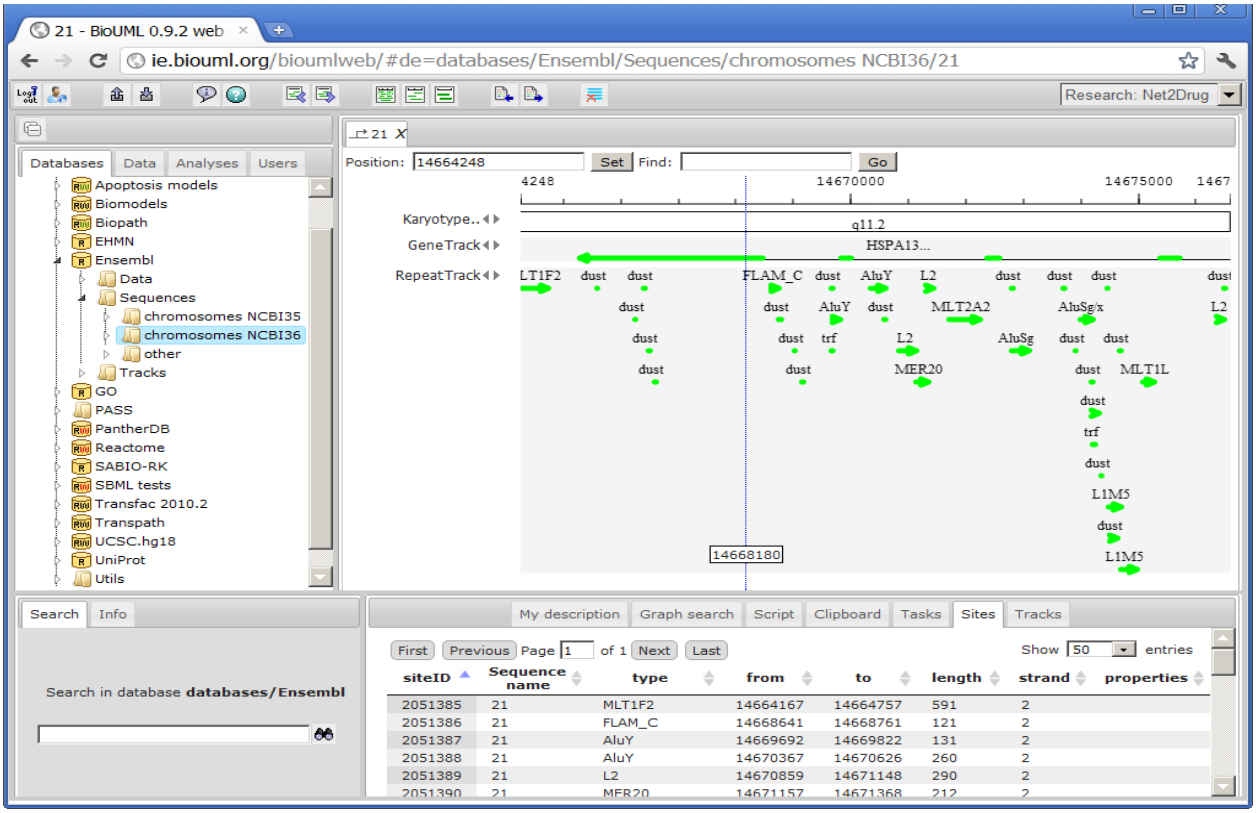


Figure 1. *BioUML web interface: top left – repository pane, top right – document pane (genome browser is opened), bottom left – search pane, bottom right – document parts (list of sites from the region selected in genome browser).*

Classification of tandemly repeated DNA families in the mouse genome

Aleksey Komissarov[1], Ekaterina Gavrilova[2], Olga Podgornaya[1]

[1]*Institute of Cytology RAS, 4 Tikhoretsky avenue, 194064, St. Petersburg, Russia, ad3002@gmail.com*
[2]*St. Petersburg State University, Universitetskaya nab. 7/9, St. Petersburg 199034, Russia*

A large portion of the mouse genome consists of tandemly repeated DNA. Here we report a genome wide analysis of the large tandem repeats with array length greater than 3 kb in the available mouse genome assemblies. Using Tandem Repeat Finder [1] and created pipeline for tandem repeats analysis, the large tandem repeats were identified in three mouse whole genome shotgun assemblies (WGSA) and two mouse genomes assemblies (reference and Celera assemblies). The analysis pipeline includes six steps: 1) filtering by array size; 2) redundancy eliminating; 3) computation distance between each pair of tandem repeat arrays; 4) creating a network of genome tandem repeats; 5) topological analysis; 6) annotating the groups found in

previous step. With this pipeline the large tandem repeats from genomes assemblies were classified into four superfamilies, eight families, and 62 annotated subfamilies including 60 that are not described yet. The main advantage of our approach is that the classification is based not only on a sequence similarity, but also on chromosome position, monomer length, array variability, and GC content.

We found only two families of previously described large tandem repeats: pericentromeric major satellite and centromeric minor satellite. The major satellite represents the main part of large tandem repeats found in WGSA and reveals high order repeat structure similar to high order repeats from primate alpha satellite. In contrast with WGSA in the genome assemblies major satellite was not assembled on chromosome except the end of chromosomes 9 and 11, although it was included in Chromosome Unknown in Celera assembly. And only WGSA and Celera's Chromosome Unknown have arrays of minor satellite.

L1-related and MTA-related families represent superfamily formed by tandemly repeated fragments of transposable elements. On the assembled chromosomes L1-related family is located preferentially within heterochromatic regions and at high concentration on the X chromosome. The superfamily of heterogeneous tandem repeats includes four families: TRPC-21A-MM, multi locus, single locus and unplaced tandem repeats. TRPC-21A-MM family resembles big human satellites.

To check genome position of three newly found families in vitro we designed short biotin-labeled probes. TRPC-21A-MM probe produces strong chromosome-specific signal to chromosomes 3 and 17. TR-22A-MM, a multi-locus family probe, hybridized to ten chromosomes. TR-54B-MM based probe gave about half of the signals on the long loops that emerge from chromosome ends.

The possibility of the existence of chromosome-specific tandem repeats had been predicted for mouse, although no reliable cytogenetic probes were available before. We report the mouse chromosome-specific (or specific to group of chromosomes) large tandem repeats found in mouse genomic assemblies. In the experimental part of work we confirmed a physical position of three families. Annotated families of large tandem repeats can be used as the library for repeats masking or as the source of cytogenetic probes for chromosome recognition and rearrangements revealing.

Search of Amino Acid Residues Crucial for Influenza Virus Hemagglutinin Anchoring Segment Organization and Interaction with Matrix M1 Protein

Anton POLYANSKY[1], Marina SEREBRYAKOVA[1], Andrei ALEXEEVSKI[2,3], Larisa KORDYUKOVA[3]

[1]*Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Russian Academy of Sciences, Moscow 117997, Russia*, [2]*Department of Bioengineering and Bioinformatic, Moscow State University, Moscow 119991, Russia* ;[3]*Belozersky Institute of Physico-Chemical Biology, Moscow State University, Moscow 119991, Russia*
kord@belozersky.msu.ru

Influenza virus belongs to Orthomyxoviridae family of enveloped viruses including type A, B and C viruses. Two major structural proteins of Influenza A virion, homotrimeric transmembrane (TM) glycoprotein hemagglutinin (HA) and matrix M1 protein underlining the viral lipid membrane from inside are thought to interact providing assembly and buddying reactions. The HA is indispensile for the delivery of the viral genome into the cell cytoplasm via fusion of the viral and endosome membranes at acid pH. Antigenic subtyping let virologists describe 16 distinct H1-H16 subtypes among type A virus falling into two major groupings, Group-1 (H1) and Group-2 (H3), while inner protein M1 is used for virus typing. The X-ray data are available for water-soluble ectodomain of various HA subtypes, and the N-terminal two thirds of M1 protein (NM-domain) probably orienting to the viral membrane.

We have applied three new approaches to characterize structural peculiarities of the HA anchoring segment (included 10 aa linker, 27 aa TM and 10-11 aa intraviral region and besides, post-translationally modified with fatty acid residues bound to three highly conserved cysteine residues) as well as its putative molecular interaction with M1 protein.

First, a unique proteolysis/mass spectrometry-based protocol allowed us to probe the HA TM domain oligomerization in micelle. We have found a surprising thing that while serine proteinase subtilisin Carlsberg cleaved Group-2 HAs exclusively in the linker region (connecting the TM and ectodomain) the Group-1 HAs were cleaved inside the TMs. Second, using molecular modeling techniques based on analysis of hydrophilic/hydrophobic properties and landscape of the TM helices we created first 3-D models of both Group-1 and Group-2 HA representatives. The predicted structures demonstrated stable behavior during MD simulations in the membrane-like environment. More compact Group-2 HA TM homotrimer complex was stabilized by several aromatic (phenylalanine) stacking pairs on the TM helix-helix interfaces. The digestion and molecular modeling data let us propose that the Group-2 HA TMs are more tightly packed in trimers compared to Group-1 ones.

Third, multiple alignments of consensus amino acid sequences of HA anchoring segments were statistically analyzed (17089 entries of HAs of type A H1-H16 subtypes and type

B HA were obtained from Influenza Research Database; http://www.fludb.org/). An N-terminal TM 10 amino acid cluster could be accentuated that exhibited extremely high conservation though all HA subtypes/types (minimal values were 98,0; 90,5; 99,4; 99,0; 99,5; 99,7; 97,9; 98,5; 99,8 and 97,5% for aa 2-11 of TM, respectively). Alternatively, C-terminal part of TM region (aa 12-27) demonstrated only 67,3% of conservation (mean minimal value for 12-27 aa of TM) albeit some specific amino acids were highly conserved within distinct subtypes. This observation implies that the C-terminal part of TM region is more variable compared to the N-terminal one. Besides, there were characteristic features specific for either Group-1 or Group-2 TMs. We speculate that the increased amino acid variability of the TM C-terminal region might be structurally adjusted to the acylation of TM utmost C-terminal conservative cysteine with various portions of stearate (C 18:0) and less hydrophobic palmitate (C 16:0). Probably, fatty acid residues help stabilize the C-terminal part of the TM homotrimer that might serve for membrane fusion successful proceeding.

This work was supported by Russian Foundation for Basic Research (09-04-01160, 10-04-91333, 10-07-00685, 11-04-91340) and by the Russian Ministry of Science and Education (MK-8439.2010.4).

Meta-analysis for discovery of disease biomarkers with implicated mechanistic models.

Ekaterina Kotelnikova[1], Maria SHKROB[1], Mikhail PYATNITSKIY[1], Nikolai DARASELIA[2]

[1]*Ariadne, Russian Federation, ekotelnikova@ariadnegenomics.com*
[2]*Ariadne, United States, nikolai@ariadnegenomics.com*

Here we propose an approach for simultaneous analysis (meta-analysis) of several microarray datasets in order to elucidate molecular pathways implicated into the specific disease and to find potential disease expression markers along with mechanistic explanations of their regulation. Duchenne muscular dystrophy (DMD) was taken as a model disease. We analyzed all available datasets from GEO database passed the samples number control and used original method based on leave-one-dataset-out validation for selection of most reproducible differentially expressed genes, regulators and pathways for further biomarkers and drug discovery. Potential expression regulators were revealed using ResNet database and Subnetwork Enrichment Analysis implemented in Pathway Studio software. Most of the selected differentially expressed genes were placed in the context of suggested regulators and pathways, resulting into the list of potential Duchenne muscular dystrophy markers with implicated mechanistic models. The proposed analysis revealed activity of several mechanisms with previously discussed role in DMD progression: muscle-related transcriptional factors (like

MYOG and MYOD1), inflammation regulators and regulators from TGFBR-SMAD pathway. We also proposed for the first time that negative expression changes in Duchenne dystrophy could be potentially explained by the members of single generalized signaling cascade (AMPK, TORC2, PPARGC1A), corresponding to the energy AMPK-TOR/insulin-glucagon-adiponectin pathway. We further speculate that the observed downregulated genes and their regulators could be responsible for the slow-fast twitch myofibers transition, the process that is known to affect the severity of DMD symptoms, making corresponding genes valuable candidates for being a potential biomarkers or drug targets.

Coexistence of different base periodicities in prokaryotic genomes as related to DNA curvature, supercoiling, and transcription

Galina Kravatskaya, Yury Kravatsky, Vladimir Chechetkin, Vladimir Tumanyan

Engelhardt Institute of Molecular Biology of Russian Academy of Sciences, Moscow, Russia, gk@eimb.ru

We performed a detailed analysis of the periodic patterns in the E. coli promoters and compared the distributions of the corresponding patterns in promoters and in the complete genome to elucidate their signal functions. The sites with the most pronounced periodic patterns appear often to be the strongest regulators that are worth being investigated experimentally. The most salient periodicities in the promoter set proved to be related to A/T composition variations, helical periodicities ~10–11 bp, spacer periodicities ~15–20 bp (termed by the distance between canonical –35 and –10 elements in the promoter region) and to three-base periodicity. Except three-base periodicity, coincident with that in the coding regions and growing stronger in the region downstream from the transcriptions start (TS), all other salient periodicities are peaked upstream of TS. We found that helical periodicities with the lengths about B-helix pitch ~10.2–10.5 bp and A-helix pitch ~10.8–11.1 bp coexist in the genomic sequences. These features were observed in both the E. coli genomic sequences and seven other complete bacterial and archaeal genomes with supercoiling of different signs. We studied also the periodicities related to the alternating purine–pyrimidine sequences potentially responsible for B-Z transition. If A- and B-like periodicities are significantly abundant in genomic DNA sequences with respect to the random ones, Z-like periodicity is underrepresented. The comparison with experimental data indicates that promoters with the most pronounced periodicities may be related to the supercoiling sensitive genes. The relaxation of negative supercoiling downregulates the genes with A- and Z-like periodicities and upregulates the genes with B-like periodicity as a form of feedback control. Besides the principal interest, the study of underlying periodic patterns may be

integrated in the general program for in silico search and in vivo assessment of the putative regulatory sites. E. coli was chosen as one of the most studied model system. Our approach is universal and may be applied to the study of genomic sequences for various organisms.

We developed a set of ad hoc programs and scripts, including two strongly optimized fourier transform programs that are based upon modern OpenMP a OpenCL technologies thus enabling us to perform fourier spectra calculations of whole bacterial genomes in a reasonable period of time (from a few hours to a few days).

This work was supported by the Molecular and Cellular Biology Program of the Presidium of the Russian Academy of Sciences.

The Just Enough Results Model (JERM) for Systems Biology Data

Olga Krebs[1], Katy Wolstencroft[2], Stuart Owen[2], Wolfgang Mueller[1], Carole Goble[2], Jacky L. Snoep[2]

[1]*Heidelberg Institute for Theoretical Studies (HITS), Germany, olga.krebs@h-its.org*
[2]*University of Manchester, United Kingdom*

The SysMO DB is a web based platform for finding, sharing and exchanging data, models and processes in Systems Biology. The SysMO DB project is developing a data access, model handling and data integration platform across the SysMO consortium (Systems Biology for MicroOrganisms), but the principles and methods employed are equally applicable to other multisite Systems Biology projects. In SysMO-DB, we adopt the minimum information model idea with our Just Enough Results Model (JERM) to link SysMO data to MIBBI standards and community ontologies.

The "Just Enough Results Model" is a minimum information model that describes the relationships between SysMO assets (i.e. experiments, data, models and publications). The JERM ("Just Enough Results Model") is the underlying model for understanding the structure and content of assets, and extracting them from their source. A JERM for any one type of data (i.e. microarray data, or metabolomic data) is the minimum data schema that the SysMO projects agree to share.

The JERM links SysMO data to MIBBI standards and community ontologies. RightField, the SysMO-DB tool for embedding ontology term selection into spreadsheets, enables JERM compliant annotation of data.

Find out more at: http://www.sysmo-db.org/

Correlating evolutionary and functional traits in vertebrates, arthropods, and fungi.

Evgenia Kriventseva

Department of Genetic Medicine and Development, University of Geneva Medical School, Geneva, Switzerland, *Evgenia.Kriventseva@unige.ch*

I would like to present the results of our resent study of relations between evolutionary and functional traits using publicly available information from knockout experiments and our orthologydata as collected in OrthoDB, freely accessible from http://cegg.unige.ch/orthodb. The study benefited from new database features: functional annotations; quantification of evolutionary divergence and relations among orthologous groups; extended phyletic profile querying, and enhanced text-based searches. Furthermore, uniform analysis across lineages as different as vertebrates, arthropods and fungi with divergence levels varying from several to hundreds of millions of years provides vital data for uncovering and quantifying long-term trends of gene evolution.

We classified 86% of over 1.36 million protein-coding genes from 40 vertebrates, 23 arthropods, and 32 fungi into orthologous groups and linked over 90% of them to Gene Ontology or InterPro annotations. Quantifying properties of orthologphyletic retention, copy-number variation, and sequence conservation, we examined correlations with gene essentiality and functional traits. More than half of vertebrate, arthropod, and fungal orthologs are universally present across each lineage. These universal orthologs are preferentially distributed in groups with almost all single-copy or all multicopy genes, and sequence evolution of the predominantly single-copy orthologous groups is markedly more constrained. Essential genes from representative model organisms, Musmusculus, Drosophila melanogaster, and Saccharomyces cerevisiae, are significantly enriched in universal orthologs within each lineage, and essential-gene-containing groups consistently exhibit greater sequence conservation than those without. This study of eukaryotic gene repertoire evolution identifies shared fundamental principles and highlights lineage-specific features, it also confirms that essential genes are highly retained and conclusively supports the "knockout-rate prediction" of stronger constraints on essential gene sequence evolution. However, the distinction between sequence conservation of single- versus multicopyorthologs is quantitatively more prominent than between orthologous groups with and without essential genes. The previously underappreciated difference in the tolerance of gene duplications and contrasting evolutionary modes of "single-copy control" versus "multicopy

license" may reflect a major evolutionary mechanism that allows extended exploration of gene sequence space.

1. Waterhouse RM, Zdobnov EM, Kriventseva EV. (2011) Correlating traits of gene retention, sequence divergence, duplicability and essentiality in vertebrates, arthropods, and fungi, *Genome BiolEvol.*, **PMID: 21148284:** 75-86.
2. Waterhouse RM, Zdobnov EM, Tegenfeldt F, Li J, Kriventseva EV. (2011) OrthoDB: the hierarchical catalog of eukaryotic orthologs in 2011, *Nucleic Acids Res.*, **PMID:20972218:** (Database issue):D283-8.

Time-sensitive inference of gene regulatory networks

Pegah Tavakkolkhah, Ralf Zimmer, Robert Küffner

Department of Informatics, Ludwig-Maximilians University, Amalienstr. 17, 80333 Munich, Germany, {pegah, zimmer, kueffner}@ifi.lmu.de

Introduction. Many algorithms were devised to deduce gene regulatory networks (GRN) from mRNA expression data. Candidate transcription factor:target gene (TF:TG) relationships are assumed more likely if the expression of the TG depends on the expression of the TF. This dependency can for instance be evaluated by Pearsons linear correlation coefficient ρ^2 [1] or by η^2 [2], a non-parametric, non-linear correlation coefficient computed from an analysis of variance (ANOVA). In particular, η^2 performed significantly better than previously published methods in the recent DREAM5 competition [3].

Inference algorithms usually neglect to analyze whether expression changes in TFs precede expression changes in TGs. We present a simple but effective approach to extend standard algorithms (exemplified by ρ^2 and η^2) by an analysis of time shifted expression patterns from time series data and report the achieved performance improvements.

Methods. Usually, interactions are ranked by a correlation c_1 (here based on ρ^2 or η^2) that relates TF levels to the corresponding TG levels measured on the same chip. We propose to compute two more correlations, c_{for} and c_{rev}. Here, a TF level at an earlier time point t_1 is also related to a TG level at a later time point t_2 yielding c_{for} (and vice versa, yielding c_{rev}). Both measurements t_1 and t_2 are derived from different time points in the same time series and satisfy the following constraint: $T_1<t_1$ & $t_2<T_2$ & $t_2-t_1>T_3$ & $T_4>t_2-t_1$, where $T_1 \ldots T_4$ are time thresholds, to be determined empirically or from biological knowledge. A time band (Figure 1, left panel) is thus defined where meaningful expression changes are expected to occur. Using $c_2= c_{\text{for}}^2 - c_{\text{rev}}^2$ we define a combined score $c=w*\text{rank}(c_1)+(1-w)*\text{rank}(c_2)$ with $w=0.9$. Candidate interactions are sorted for relevance according to c. We evaluate c by a *directionality* test distinguishing known interactions TF:TG (true, 51%) from their reverse TG:TF (false, 49%) as well as by an *inference*

test distinguishing known interactions (true, ≈1%) from all other possible interactions (false, ≈99%). In the directionality test, #true is larger than #false due to bidirectional interactions.

Expression datasets, known TF-TG relationships and evaluation protocols, e.g. area under precision-recall curve (AUPR), were used as in the DREAM5 assessment [4].

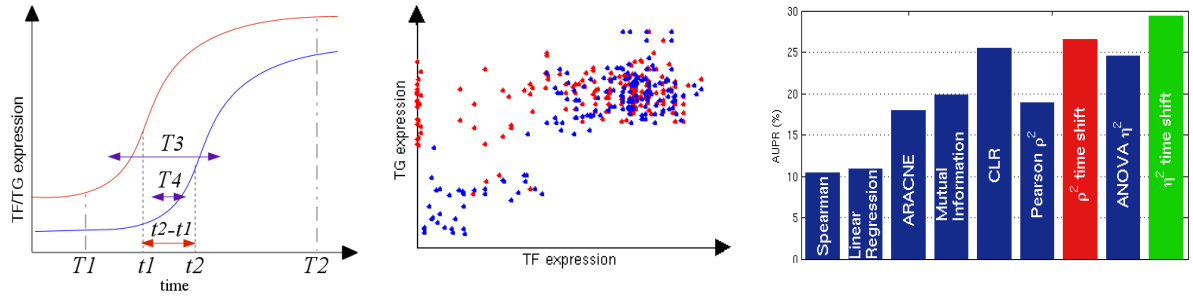


Figure 1. Expression of target genes (blue, left panel) lags behind the expression of their regulators (red). Thus, if earlier time points of the TF are correlated to later time points of the TG (blue, middle panel) the correlation between TF and TG will be higher than in the inverse case (red). Existing methods can be improved by incorporating an analysis of time delays (red and green, right panel).

Results and Discussion. TF mRNA needs to be exported from the nucleus, translated into TF protein, which has to be imported back into the nucleus before expression changes of the TF can become effective. This leads to considerable time delays between TF and TG expression changes. The majority of current inference methods neglect a dedicated analysis of time series but solely focus on correlation for the inference of causal dependencies. Testing for such time delays (i.e. TF expression changes preceding TG expression changes) should therefore improve the accuracy of the network inference algorithms. We determined that such temporal information can be extracted from expression data by a directionality test that resulted in AUPR of 81.9% on DREAM5 artificial data. We showed that different methods can be improved by integrating temporal dependencies, e.g. Pearsons correlation ρ^2 from 18.9 to 26.5% AUPR and ANOVA's η^2 from 24.5 to 29.3% AUPR. Other commonly used approaches, e.g. based on mutual information, should also benefit from time series analysis. The presented method is very simple and we expect additional gains in performance by using more involved analyses of time series data.

1. Butte, A. and Kohane, I. (1999). *Proc AMIA Symp*, 711-715
2. Cohen, J. (1973). *Educ Psychol Meas*, **33(1):** 107-112.
3. R. Küffner et al. (2011) Inferring Gene Regulatory Networks by ANOVA, *submitted to ISMB2011*.
4. DREAM5 setting and data: http://wiki.c2b2.columbia.edu/dream/index.php/D5c4.

Preferred Pair Distance Templates for Analysis of Transcription Regulation Code

Ivan V. Kulakoskiy[1,2], Alexander A. Belostotsky[2], Artem S. Kasianov[1], Yulia A. Medvedeva[2,3], Irina A. Eliseeva[4], Vsevolod Makeev[2,3]

[1]*Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Russia* ivan.kulakovskiy@gmail.com
[2]*Research Institute for Genetics and Selection of Industrial Microorganisms, Russian Federation*
[3]*Vavilov Institute of General Genetics, Russian Academy of Sciences, Russian Federation*
[4]*Institute of Protein Research, Russian Academy of Sciences, Russian Federation*

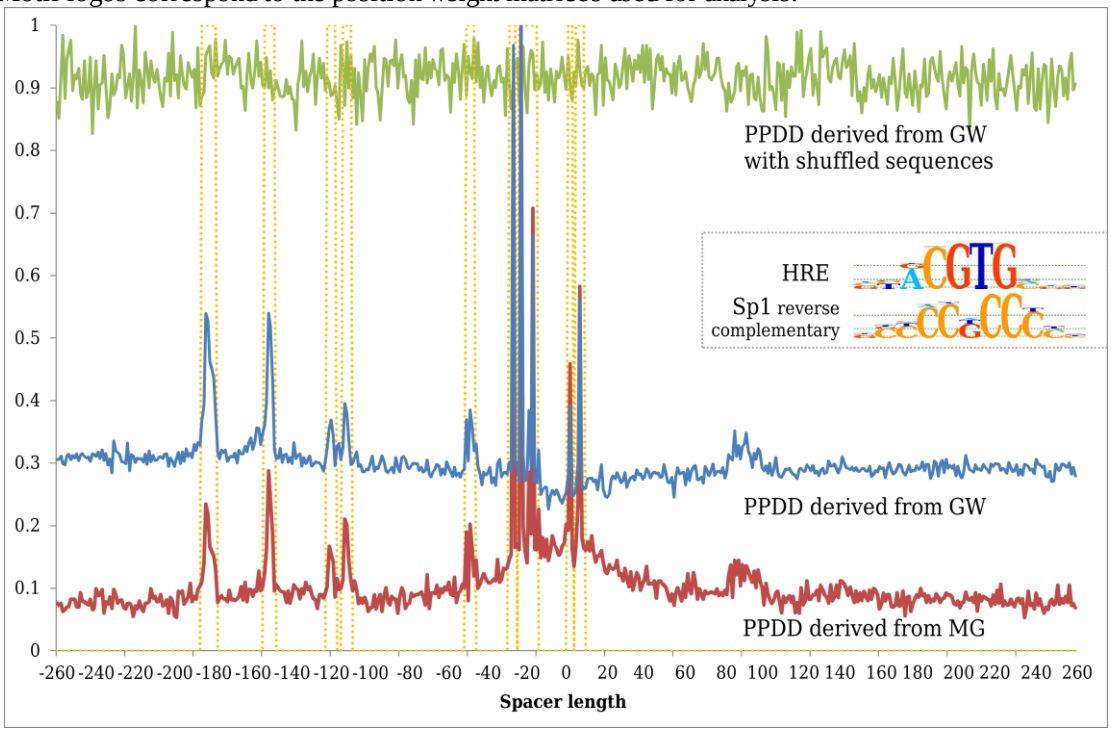
Motivation and Aim. The gene expression control and the corresponding regulatory code in higher eukaryotes remain unclear. It is a difficult task to understand how a one-dimensional regulatory DNA text consisting of many binding sites directs formation of huge protein complex that controls gene expression in a particular tissue or in specific conditions. Some ideas are given by the classic concept of "composite elements" consisting of binding sites for different transcription factors (TFs) separated by specific distances [1]. Information about possible scale of the distances and their specificity remains insufficient despite more than 15 years of study. Recently it was clearly shown [2], [3] that some pairs of transcription factor binding sites (TFBS) have preferences to more than one distance of mutual localization. We show that distance preferences themselves are extremely well exhibited and this phenomenon appears to be much more common at least in the case of Homo sapiens TFs involved into regulation of response for hypoxic conditions [4].

Materials and Methods. We construct the preferred pair distance distributions (PPDD) for the pairs consisting of HRE (the hypoxia responsive element, the binding site for the HIF-1α:ARNT dimer) and binding sites for other hypoxia-related transcription factors having known protein-protein interaction with the HIF-1α. We use the set of genome-wide promoter regulatory regions (the GW set) extracted using the UCSC hg18 annotation. The similar set (the MG set) is made from GW by masking exons and repetitive regions. To identify putative binding sites we use the well-known positional weight matrix model built from various pregenomic and ChIP-Seq data by ChIPMunk [5] tool. To construct the Preferred Pair Distance Distribution we count the number of regulatory regions (i.e. the number of genes) having a selected binding site pair in a given orientation separated by a given spacer. We then extract the set of well-exhibited "peaks" and call it the Preferred Pair Distance Templates (PPDT) which define the small subset of "possibly functional" distances. The Figure 1 shows an example of PPDD/T and its robustness versus the genome-masking procedure.

Results and Discussion. Preferred distances between TFBS seem to be related either to the direct interaction between TFs or to the indirect interaction via adapter proteins or particular

chromatin structures. The PPDD/T concept makes it possible to discriminate the set hypoxia-dependent regulatory regions from the GW set. Moreover it can be used to predict novel regulatory regions and target genes as well as TF interactions and possible regulatory complexes. We believe that similar patterns can be observed and used for other regulatory systems and that further understanding of PPDD/T should allow deeper understanding of a regulatory code in higher eukaryotes.

Figure 1. The PPDD and PPDT peaks for the HRE-reverse complementary Sp1 binding site pair. PPDDs from three different sequence sets are shown; the curves are normalized for their maximum. The Y axis corresponds to the fraction of sequences in the set having a pair of sites separated by the selected spacer (at X axis). HRE is located at zero X. Motif logos correspond to the position weight matrices used for analysis.



Acknowledgments and Funding. We thank Biobase GmbH and personally Alexander Kel for granting us the access to the TRANSFAC release 2010.1 and Dmitrijs Lvovs, Valentina Boeva and Adam Arents for valuable comments. This study was supported by the Presidium of the Russian Academy of Sciences program in Cellular and Molecular Biology.

1. V. Matys et al. (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes, *Nucleic Acids Res*, **1:**34(Database issue):D108-10.
2. K.D. Yokoyama et al. (2009) Measuring spatial preferences at fine-scale resolution identifies known and novel cis-regulatory element candidates and functional motif-pair relationships, *Nucleic Acids Res*, **37(13):**e92.
3. V. Shelest et al. (2010) DistanceScan: a tool for promoter modeling, *Bioinformatics*, **26(11):**1460-2.
4. I.V. Kulakovskiy et al. (2011) Preferred distances between transcription factor binding sites, *Biophysics*, **56(1):**114-6.
5. I.V. Kulakovskiy et. al. (2010) Deep and wide digging for binding motifs in ChIP-Seq data, *Bioinformatics*, **26(20):**2622-3.

Use of natural compounds from plant sources as AchE inhibitors for the treatment of early stage Alzheimer's Disease an insilico approach

Amrendar Kumar, Abhilasha Singh
Amity Institute of Biotechnology, Lucknow campus, UP, India, amrendar2290@gmail.com

Traditionally, drugs were discovered by testing compounds manufactured in time consuming multi-step processes against a battery of in vivo biological screens. Promising compounds were then further studied in development, where their pharmacokinetic properties, metabolism and potential toxicity were investigated. Here we present a study on herbal lead compounds and their potential binding affinity to the effectors molecules of major disease like Alzheimer's disease. Clinical studies demonstrate a positive correlation between the extent of Acetyl cholinesterase enzyme and Alzheimer's disease. Therefore, identification of effective, well-tolerated acetyl cholinesterase represents a rational chemo preventive strategy. This study has investigated the effects of naturally occurring nonprotein compounds polygala and bulbocapnine that inhibits acetylcholinesterase enzyme. The results reveal that these compounds use less energy to bind to acetylcholinesterase enzyme and inhibit its activity. Their high ligand binding affinity to acetylcholinesterase enzyme introduce the prospect for their use in chemopreventive applications in addition they are freely available natural compounds that can be safely used to prevent Alzheimer's Disease.

Comparative analysis of metabolic profiles for human gut microbiota using ABI SOLiD sequencing

Irina Kuncevich, Alexander Tyakht
Institute for Physico-Chemical Medicine, Russian Federation, irakuncevich@gmail.com

Human gut houses hundreds of various species of microbes, mainly bacteria. These organisms are interconnected into a united ecosystem (microbiota) creating interactive network between them and human organism as well. With the advent of next-generation high-throughput DNA sequencing technologies, a deeper insight in its phylogenetic and functional composition has become feasible. Genomic sequences for a wide range of bacteria composing microbiota can be assembled. It is necessary to design efficient methods for metagenomic analysis of such vast amount of data — particularly, for statistical comparison between separate subjects and groups.

Many of microbes in microbiota are poorly examined and are unclassified. It is common that genomes of separate close species have similar composition therefore assembly from short DNA reads is ambiguous. In order to avoid problems associated with precise taxonomic characteristics of metagenomic information, we propose a unified analysis of microbiota based on gene-centric view on its metabolic functions. The analysis pipeline is based on hydrid assembly combining assembly de novo and mapping to reference genomes.

For the aims of functional analysis, the longer fragments assembled from short reads should be at least long enough to find open-reading frames (ORFs) and predict proteins. Assembly de novo can hardly provide such quality when a metagenomic sample contains hundreds of diverse species. An alternative way to obtain long contigs from short reads depends on a priori information about phylogenetic composition of the metagenomic sample. With the help of sequence alignment software, the reads are aligned to a reference set of genomes which are supposed to be typically abundant in similar environments (i.e., human intestine). Coverage of reference genomes produced by mapping the reads allows to estimate metagenomic composition both qualitatively and quantitatively. Afterwards, the remainder can be assembled de novo.

The reconstructed DNA sequences are searched for open-reading frames and genes corresponding to metabolic proteins. The set of such enzymes can be reconstructed into metabolic pathways of microbes [3]. An experimental approach to pathway analysis is suggested: due to close ecological interconnection between organisms comprising microbiota, separate pathways can be combined into a total metabolic network. Such network can be

extended by adding quantitative information — i.e., weights given to its sides corresponding to the abundance level for such connection in the system. The structure can be characterized using mathematical methods from graph theory. Particularly, functionally important subnetworks can be selected.

The results of analysis provide the way to suggest hypotheses concerning the causes of physiological failures in human health related to unbalanced metabolic potential of gut microbiota. The pipeline for batch analysis of metagenomic data is demonstrated by example of several metagenomic datasets. The DNA samples have been taken from Russian citizen, both healthy and having certain problems with gut health, in the course of Russian metagenomic research initiative which includes collection of metagenomic samples by a number of medical insitutions in Russia and DNA sequencing followed by bioinformatic research. One of the tasks of analysis is examining the datasets for features specific to various social groups and citizen of Russia in comparision with available data from Europe and other countries, i.e. Human Microbiome Project [1] and MetaHit [2].

Investigation of NAD^+ binding to glyceraldehyde 3-phosphate dehydrogenase

Mikhail L. KURAVSKY, Elena V. SCHMALHAUSEN, Vladimir I. MURONETZ

Lomonosov Moscow State University, Leninskie Gori, Moscow, 119991, Russia, mkuravsky@gmail.com

Glyceraldehyde 3-phosphate dehydrogenase (GAPDH) is a homotetrameric glycolytic enzyme catalysing oxidative phosphorylation of glyceraldehyde 3-phosphate to glycerate 1,3-bisphosphate coupled with reduction of NAD^+ to NADH. Humans possess two homologous isoenzymes of GAPDH with 68% identical sequences: somatic (GAPDH-1) and testis-specific (GAPDH-2). GAPDH-1 is a well-studied enzyme which is present in all cells. It was discovered to exhibit negative cooperativity: the affinity for NAD^+ lowers with the amount of NAD^+ already bound [1]. GAPDH-2 is found only in sperm tails. Recent studies by our group established that GAPDH-2 is also present in certain malignant tumors, *viz.* melanomas (data to be published).

In this work, we apply both experimental and bioinformatical approaches to investigate NAD^+ binding characteristics of human GAPDH isoenzymes. This does not only provide new data on little-studied GAPDH-2, but also give some insight into the interactions resulting in negative cooperativity.

The NAD^+ binding constants of human GAPDH isoenzymes were estimated by means of fluorescence quenching titrations. GAPDH-2 was found to exhibit higher affinity for NAD^+ and weaker negative cooperativity than GAPDH-1. This peculiarity seems for us to be an adaptation to the spermatozoa metabolism. In contrast to the majority of human cells, spermatozoa are known to generate energy mainly in the course of glycolysis [2]. It leads to the necessity in glycolytic enzymes capable of working in low NAD^+/NADH conditions such as GAPDH-2. It is significant that malignant tumors expressing GAPDH-2 are also known to utilize glycolysis as the main energy source.

The limitations of experimental methods forced us to switch to bioinformatics for more detailed study of interactions between GAPDH subunits. Ligand-protein affinities were estimated from molecular dynamics simulations by means of linear interaction energy (LIE) method [3]. The NAD^+ binding free energies exhibited by each of 4 binding sites in GAPDH tetramers were calculated to be -3 kJ/mol, -7 kJ/mol, -33 kJ/mol and -60 kJ/mol for GAPDH-1, as well as -18 kJ/mol, -37 kJ/mol, -47 kJ/mol and -54 kJ/mol for GAPDH-2. These values meet the experimental ones confirming the model quality. Further analysis of human GAPDH-1 crystal structure (PDB ID 1znq) revealed that each of 4 subunits (usually designated as O, P, Q and R) makes contacts with 2 other subunits, *e.g.* subunit O makes contacts with subunits P and R. Therefore, negative cooperativity should be mediated by either OP-type or OR-type interactions. To distinguish between these two possibilities, the affinities for NAD^+ were estimated for both OP and OR dimers. OR dimer was discovered to retain negative cooperativity (-24 kJ/mol and -56 kJ/mol), while OP dimer was not (-43 kJ/mol and -54 kJ/mol). Free energy of interactions between O and R subunits (-348 kJ/mol) was calculated to be 3 times greater than between O and P subunits (-105 kJ/mol). These findings mean that GAPDH tetramer may be considered as a dimer of dimers (OR + PQ) and negative cooperativity of NAD^+ binding is mediated by interplay within OR and PQ dimers.

The work was supported by Russian Foundation for Basic Researches (grants 09-04-01122-a and 09-04-92740-NNIOM_a). Molecular dynamics simulations were performed at SKIF-MGU “Chebyshev” supercomputer.

1. J.E.Bell, K.Dalziel (1974) Studies of coenzyme binding to rabbit muscle glyceraldehyde-3-phosphate dehydrogenase, *Biochim Biophys Acta*., **391:**249–258.
2. C.Mukai, M.Okuno (2004) Glycolysis plays a major role for adenosine triphosphate supplementation in mouse sperm flagellar movement, *Biol Reprod*., **71:**540–547.
3. J.Aqvist et al. (2002) Ligand binding affinities from MD simulations, *Acc Chem Res*., **35:**358-365.

BioUML: a plug-in for model reduction

Elena Kutumova[1], Andrei Zinovyev[2], Ruslan Sharipov[3]

[1] *Institute of Systems Biology, Design Technological Institute of Digital Techniques SB RAS, Novosibirsk, Russia, e.o.kutumova@gmail.com*
[2] *Institute Curie, Paris, France, andrei.zinovyev@curie.fr*
[3] *Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia, shrus79@gmail.com*

BioUML (http://www.biouml.org) is an open source integrated Java-based platform for systems biology. Plug-in based architecture makes it flexible allowing to extend existing and add new functionality (e.g., for modeling and simulation of complex living systems). To provide extended possibilities for complicated model optimization we developed a plug-in including the methods for model reduction. The toolbox will implement a series of classical and novel methods for finding asymptotic solutions for the systems of ordinary differential equations appearing in chemical kinetics such as quasi steady-state, quasi-equilibrium, lumping approaches, finding limiting systems and low-dimensional dynamics approximations [1].

The plug-in already includes the following methods:

1. analysis of linear, monomolecular and pseudo-monomolecular reactions, where a linear reaction is characterized by the linear kinetic law, a monomolecular reaction has only one reactant or modifier, and a reaction is considered as pseudo-monomolecular if ratios of its reactants/modifiers are greater than a specified threshold or smaller than a threshold^{-1};
2. steady-state analysis;
3. quasi-steady state analysis detecting particular fast intermediates which amount is much smaller than amounts of the basic components;
4. stoichiometric analysis deriving the linear relationship of the model between the flux rates of the (enzymatic) reactions and the derivatives of the reactant (enzyme) concentrations;
5. flux rates analysis.

Some additional methods including mass conservation analysis and the method of species dynamics approximations based on the principal component analysis are under development now.

Using the methods of the model reduction plug-in, we considered a model of the apoptotic machinery regulation [2]. Reduction of this model permitted to decrease its dimension and detect the key parameters defining its dynamics based on experimental data.

1. A. N. Gorban, O. Radulescu, A. Y. Zinovyev (2009) Asymptotology of Chemical Reaction Networks. *Chem Eng Sci*., **65**:2310-2324.
2. M. Bentele, et al. (2004) Mathematical modeling reveals threshold mechanism in CD95-induced apoptosis. *The Journal of Cell Biology*, **166**:6.

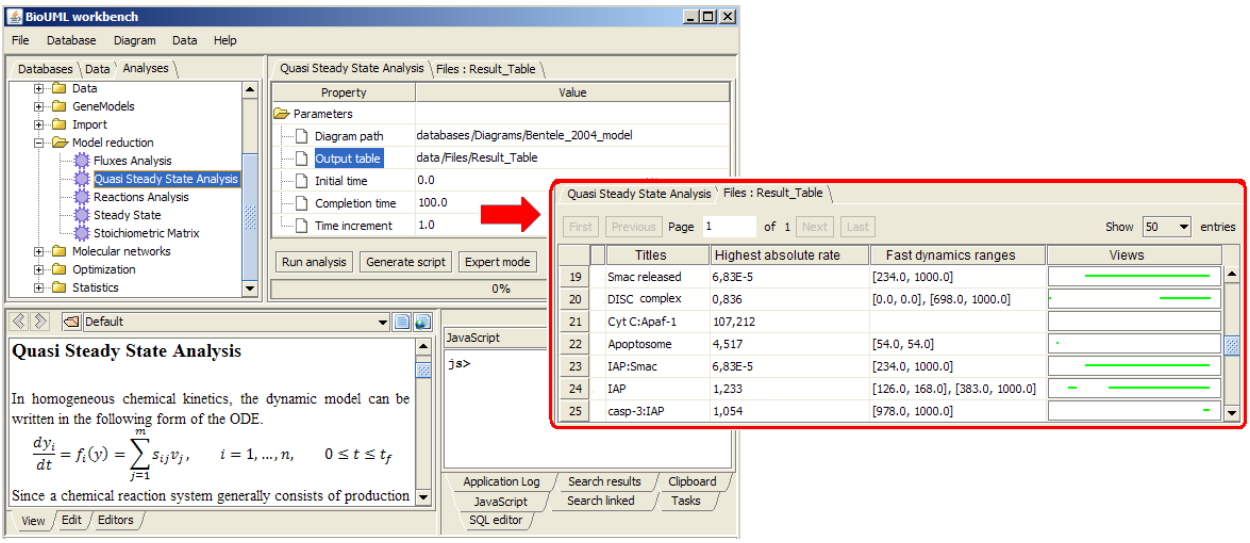


Figure 1. A user interface of the model reduction plug-in provided by BioUML workbench. The upper left panel includes the list of available methods. The lower panel represents description of the Quasi-Steady State method. The right panel contains the method parameters. The result of analysis is represented as a table including columns with species titles, values of the highest absolute rates within the simulation time interval, text and visual representation of fast dynamics ranges.

Ovarian Cancer Patient's Risk Stratification Based on miRNA-mRNA Interctome

Vladimir Kuznetsov[1], Tang Zhiqun[1], Efthimios Motakis[1],
Jean Paul Thiery[2], Anna Ivshina[2]

[1] *Bioinformartics Institute, Singapore, vladimirk@bii.a-star.edu.sg*
[2] *Institute of Molecular and Cellular Biology, Singapore*

We examine the associations of microRNA (miRNA), their host genes and their target genes with survival history and treatment protocols of 494 epithelial ovarian cancer (EOC) patients. We design a novel data-driven risk assessment strategy (DRAS) to assign cancer patients into two or more classes from miRNA-mRNA molecular fingerprints associating with patient treatment protocols and assessing survival information. Feature selection by DRAS is based on modification of the Data-Driven Grouping model (1). DRAS identifies the optimal partition of EOC patients into low, medium and high risk groups of the disease progression.

Using DRAS, we found 81 miRNAs which expression pattern is strongly associated with EOC patient survival. 40 of these survival significant miRNAs form tightly-connected interactive gene networks including several key cancer-related and stem cells associated protein-coding genes and shedding a light on a role of the miRNA-RNA interconnections in the control

of EOC. Less than 50% of these miRNAs have been discussed in the association with the human ovarian cancer studies. We developed a novel statistically weighted voting algorithm for partition of the patients according their survival data. This method was implemented to determine the robust patient's partition into three groups, and selected 50 miRNA-host gene and miRNA-target gene pairs. This interactive pairs consist of the synergistic survival signature revealing three patient' patterns well correlated with metastasis events, disease recurrence and chemotherapy treatment response. Our disease risk-association signature stratifies the patients with strong confidence level which have been not reached via application of the conventional clinical biomarkers (Figure 1).

Acknowledgements. The authors would like to acknowledge Agency for Science, Technology and Research (A*STAR) for funding this study.

1. E. Motakis, A.V. Ivshina AV, V.A. Kuznetsov (2009) Data-driven approach to predict survival of cancer patients, *IEEE Eng Med Biol Magasine* , **28**(4): 58-66.

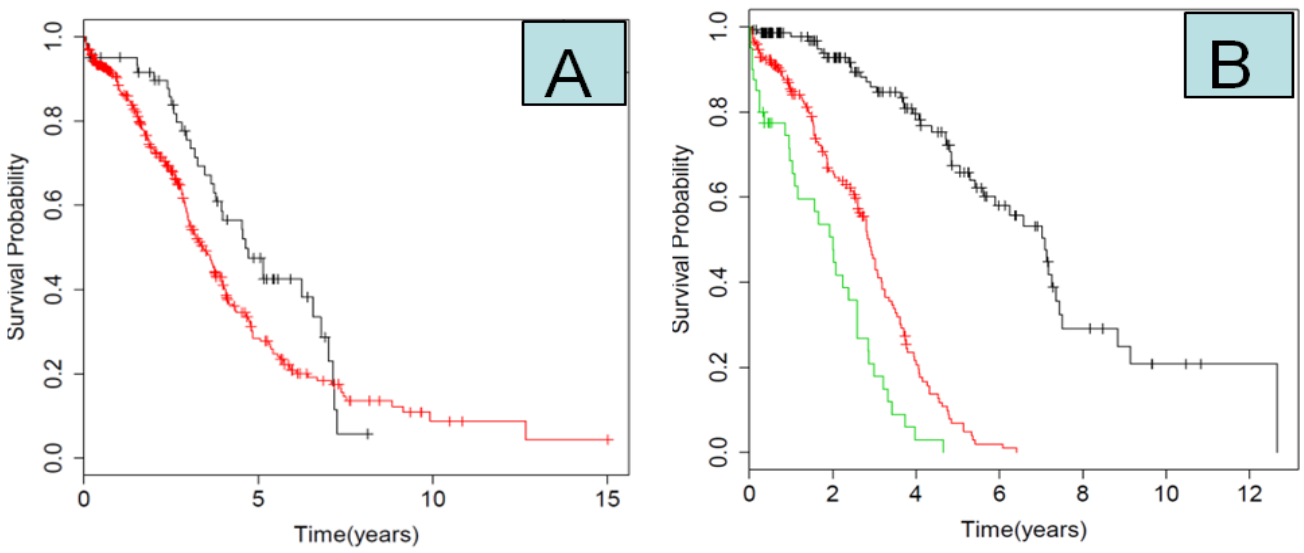


Figure 1. The Kaplan-Meier survival curves of EOC patient defined by (A) histo-pathological indicators (grades 1-2 vs grades 3-4) and (B) 50 miRNA-mRNA classifier.

BioUML: the PASS Plug-in for Prediction of Biological Activity of Substances on the Basis of their Structural Formulae

A.V. ZAKHAROV, D.A. FILIMONOV, A.A. LAGUNIN, V.V. POROIKOV

Institute of Biomedical Chemistry of RAM S, Moscow, Russia
vladimir.poroikov@ibmc.msk.ru

N.I. TOLSTYKH, F.A. KOLPAKOV

Institute of Systems Biology, Design Technological Institute of Digital Techniques SB RAS, Novosibirsk, Russia,
fedor@biouml.org

Computer system PASS (Prediction of Activity Spectra for Substances) is able to predict biological activities (pharmacological main and side effects, mechanisms of action, etc) of a substance using information about its structural formula (http://www.pharmaexpert.ru). Prediction by PASS is based on the SAR (structure-activity relationships) analysis of the training set containing more than 250,000 compounds which have more than 4000 types of biological activity (1, 2). The mean accuracy of PASS prediction calculated by leave-one-out cross-validation is about 95%. To provide the best quality of prediction, new information about biologically active compounds is collected permanently from papers and electronic sources and, after the expert evaluation, is regularly added to the training set.

BioUML (Biological Universal Modeling Language; http://www.biouml.org) is a Java-based platform for support of systems biology-related investigations. This platform has a plugin-based architecture that allows adding new modules with methods and now combines a vast set of different methods for comprehensive analysis, formal description and mathematical modeling of living systems. To integrate possibilities of the biological activity predicting system in BioUML a PASS plugin was developed and implemented both for stand-alone and web versions.

The PASS plugin uses MOL format data (http://www.mdli.com) as input and the results of prediction are generated as tables and stored in a database. Figure 1 represents a result of such prediction for a set of substances displayed in the BioUML web user interface. Optimized prediction procedure allows obtaining results even for large chemical databases not using hi-end hardware configuration. The developed plugin is dedicated to provide a tight junction between the identification of drug targets and design/search of prospective drugs on the basis of data accumulated in pharmacology-related databases.

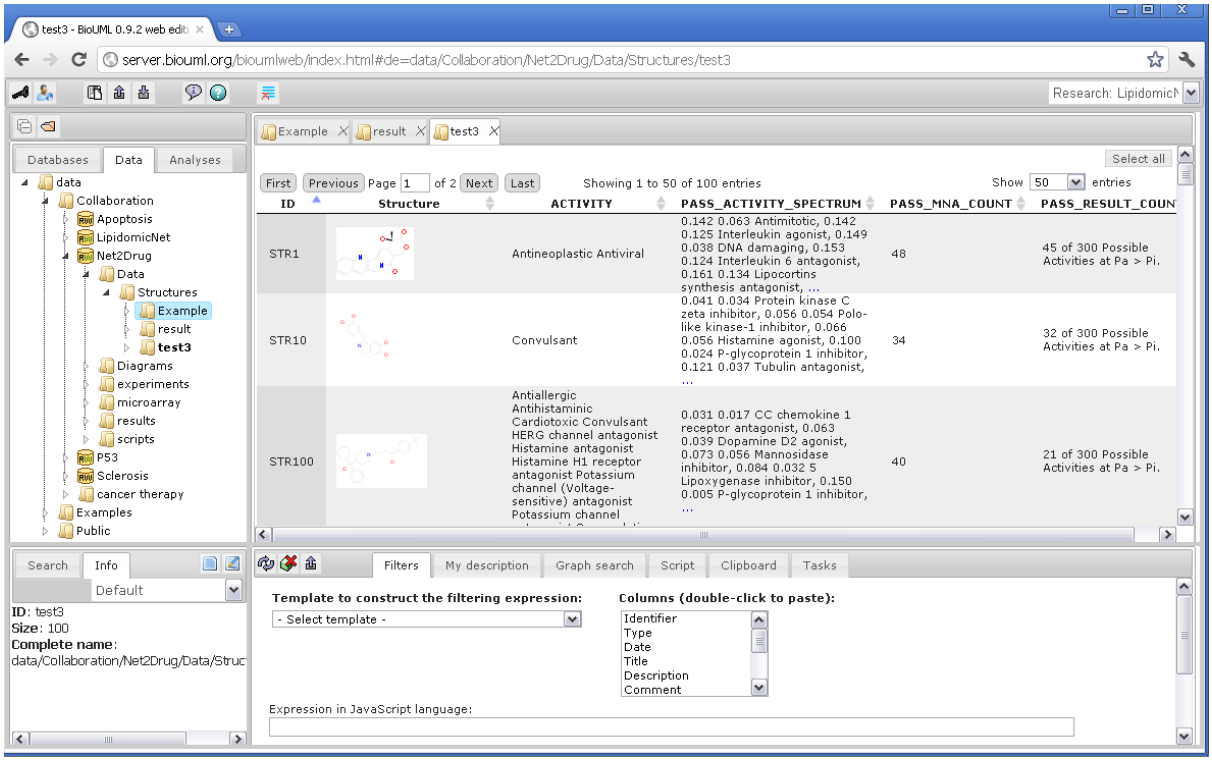


Figure 1. *BioUML web interface, top right pane - biological activities predicted using the PASS plug-in for a set of substances.*

1. D.A. Filimonov and V.V. Poroikov (2008) Probabilistic approach in activity prediction. In: Chemoinformatics Approaches to Virtual Screening. Eds. Alexandre Varnek and Alexander Tropsha. Cambridge (UK): RSC Publishing, pp.182-216.
2. A. Lagunin et al. (2010) Multi-targeted natural products evaluation based on biological activity prediction with PASS, *Curr Pharm Des*. **16**:1703-1717.

Modeling of phage infection in prokaryotic communities by Evolutionary Constructor program

Sergey Lashin, Valentin Suslov, Yury Matushkin

Institute of Cytology and Genetics SB RAS, Lavrentiev av. 10, Novosibirsk, Russia, lashin@bionet.nsc.ru
Novosibirsk State University, Pirogova 2, Novosibirsk, Russia

Prokaryotes living in natural habitats usually form communities, where various species are integrated together both in food (metabolites exchange and symbiosis) and genetic relations. In such communities, phages act as egoistic DNA, horizontal gene transfer vectors. They also regulate cells' lysis, expansion of egoistic DNA (abortive infection) and genes transfer. Although the processes of coevolution in the "phage-host" system are well-studied on the basis of population models, the coevolution between phages and symbiotrophic ecosystems have not been modeled well. In order to provide such type of modeling we have modified our program "Evolutionary constructor" (available at http://evol-constructor.bionet.nsc.ru/ ; Lashin et al, 2007, 2010): new types of objects were added – *phage populations*, and *infected cells populations*. Infection process includes the following stages: infestation of cells, intracellular phages reproduction, and phages transport into environment by cells lysis. Reproduction cycle of infected cells may also be lysogenic. In this case no phages are transported into environment, and prophages arise. The genetic spectra arithmetic allows us to consider both lytic and lysogenic cycles in terms of one polymorphic population. Moreover, the phages' polymorphism, origin of novel strains (via mutations), and competition between strains are also described using this arithmetic.

We have studied several models of interactions between a polymorphic bacterial population and a phages population. Phages strains possessing various abundance (copy-number of phages being produced at the moment of cell lysis) were considered. It has been shown that the infection of a population by low-copy-number phages leads to a splitting of a population into two subpopulations – healthy and infected, the sizes of which become stable after a certain period of time (fig. 1a). Contrariwise, the infection of a population by high-copy-number phages leads to the origin of oscillatory regimes and the death of a system in prospect (fig. 1b). The dynamics of phages populations of various copy-numbers is shown on the fig. 2. We have also studied models of interactions between bacterial communities and phages population. It has been shown, that in conditions promoting genome amplification for community members, the moment of infection plays significant role for further possible regimes of functioning.

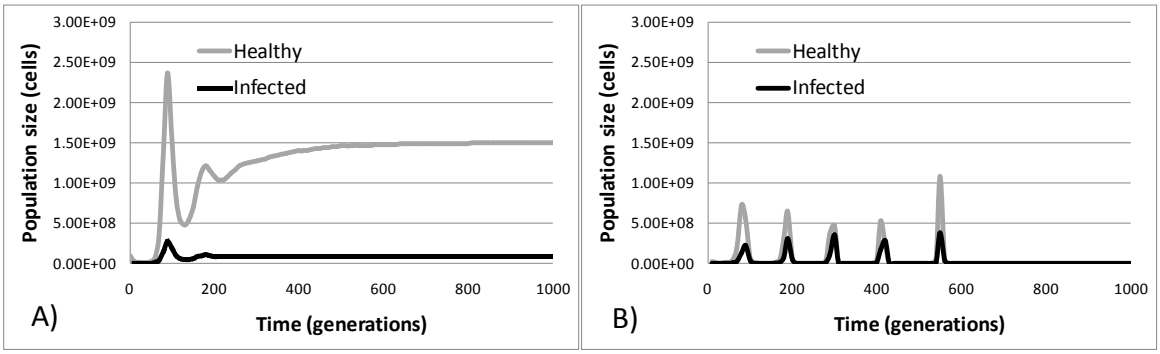


Fig.1. Dynamics of bacterial populations infected by phages: A) low-copy-number; B) high-copy-number.

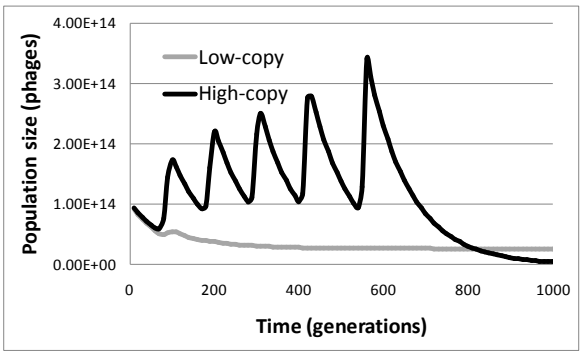


Fig.2. Dynamics of phages populations (high- and low-copy number).

Acknowledgments to the grants: RFBR (10-04-01310), RAS Program “Biosphere origin and evolution of geo-biological systems”, Program “Molecular and cellular biology” (A.II.6).

1. S.A. Lashin et al. (2007) Simulation of coevolution in community by using the "Evolutionary Constructor" program. *In Silico Biology*, **7**: 261-275.
2. S.A. Lashin et al. (2010) Comparative modeling of coevolution in communities of unicellular organisms: adaptability and biodiversity. *J Bioinf and Comp Biol*, **8**: 627-643.

Mutational load generates genomic modularity

Ralf Bundschuh[1], Juliette de Meaux[2], Michael Lassig[3]

[1]*Ohio State University, United States*
[2]*University of Muenster, Germany*
[3]*University of Cologne, Germany, mlaessig@uni-koeln.de*

Complex molecular traits often evolve under significant mutational load, even in organisms with low point mutation rate. These systems pose an old, but unresolved question: Does mutational load cause robustness against deleterious mutations? Here, we address this question for processing of miRNAs, a molecular function that is universal across plant genomes. The processing mechanism uses an extended stem in the pre-miRNA secondary structure. We show that mutational load generates genomic modularity: it shortens the DNA sequence segment encoding this stem. Moreover, mutational load directly affects the stem phenotype: it leads to more compact stem configurations with an increased average binding energy per base pair. Our analysis is based on genomic sequence data of pre-miRNAs in Arabidopsis. From these data, we infer an empirical, biophysically grounded fitness landscape for mi-RNA processing, and we compare the evolution of populations with different mutation rates in this landscape. A theoretical model of mutational load explains the emergence of genomic modularity as a generic feature of quantitative traits evolving in a complex fitness landscape. Genomic modularity, in turn, generates mutational robustness and facilitates the independent evolution of different functions. We conclude that mutational load has direct effects on genomic architecture and molecular functions.

Mathematical Approach to account for mutations in bladder tumours

CALZONE Laurence[1], CHAOUIYA Claudine[2], REMY Elisabeth[3], RADVANYI François[4]

[1]*INSERM U900 – Institut Curie – Mines ParisTech, France, France, laurence.calzone@curie.fr*
[2]*TAGC, INSERM U 928 Marseille, Instituto Gulbenkian de Ciência, Portugal*
[3]*Institut de Mathématiques de Luminy, France*
[4]*Équipe Oncologie Moléculaire, UMR 144, CNRS, France*

The E2F transcription factors are key regulators of the cell cycle. The E2F family members are categorized in two sections, the activators (E2F1, E2F2, E2F3a) and the inhibitors (E2F3b, E2F4, E2F5, E2F6, E2F7, E2F8) of proliferation. Most of these transcription factors are inhibited by the pocket proteins, RB and p107, during G0 and G1 phases and are released at the G1/S transition by phosphorylation and inactivation of the RB-like proteins. Among the E2F family genes, E2F3 has raised a lot of interest since it is subject to amplification in various cancers such as bladder, lung or prostate cancers. The E2F3 locus gene encodes the two isoforms E2F3a and E2F3b, which show both complementary and antagonist behaviours according to cell context.

Using bioinformatics and systems biology approaches, we investigated the role of E2F3 in tumour cells and their differences with E2F1 with which E2F3 is believed to share many functions. We built two networks, one descriptive reaction network showing the links between RB and p53 pathways and the involvement of the E2F1 and E2F3 transcription factors and another one, an influence network derived from the first reaction map. The reaction network includes 56 species (14 proteins, 14 mRNA, 13 genes), 4 inputs (DNA damage, growth factors, FGFR3, and TGFb) and 64 reactions. The corresponding influence network was reduced to 14 species, 4 inputs and 3 outputs (apoptosis, proliferation and growth arrest). A mathematical model based on the logical formalism was translated from the network. The dynamical model accounts for diverse phenotypes of both normal and mutant cells found in the literature in response to growth and DNA damage signals. Moreover, it allows to verify the hypotheses on how a cell becomes invasive in bladder tumours, through which signalling pathways and with which type of alterations or mutations invasiveness is associated. We also perform explorative studies on expression data for each individual gene and on possible correlations for different pairs of genes. This work is in progress and the analysis is still exploratory.

The Power of Complex Trait Rare Variant Association Methods

Suzanne M. Leal

Department of Molecular and Human Genetics, Baylor College of Medicine, One Baylor Plaza, 700D, Houston, TX 77030, USA; suzannemleal@gmail.com

There is currently great interest in detecting rare variant associations using next generation sequence data. A large number of association methods which aggregate variants across a region e.g. a gene have been developed specifically to analyze rare variant data. It is not clear which existing method is the most powerful and should be applied to test for associations using exome/genome sequence data. It is difficult to compare rare variant association methods because there is no standard to generate data and often the comparisons are biased. To fairly compare rare variant association methods it is necessary to generate data using realistic population demographic and phenotype models. The power was compared for 12 methods to detect associations for qualitative and quantitative traits. Power was evaluated for case-control, extreme quantitative trait sampling and population based study designs. For each method, power was determined for scenarios which included 1. analysis of a. only rare variants & b. rare (<1%) and low (1-5%) frequency variants; 2. detrimental and protective variants within a gene region; 3. misclassification a. exclusion of causal variants & b. inclusion of non-causal variants; 4. different underlying population demographic model for both Africans and Europeans and 5. gene size. It was observed that there is not a single method that is most powerful in all situations. The majority of rare variants methods had only small incremental difference in power. Rare variant association methods which were powerful in a variety of situations include the Variable Threshold (VT) method, Weighted Sum Statistic (WSS) and Kernel Based Adaptive Cluster (KBAC) method. Those methods which were developed to detect associations when both protective and detrimental variants are within an associated region (e.g. C-alpha) are usually less powerful than more general rare variant association methods. The evaluated methods also vary in their computation efficiency and ability to control for confounders.

Adaptive amino acid replacements triggered by indels in Drosophila proteins

Evgeniy Leushkin, Georgii Bazykin, Alexey Kondrashov
M. V. Lomonosov Moscow State University, Russian Federation, leushkin@gmail.com

The maps relating amino acid sequences to fitness, the fitness landscapes, have a complex structure which remains poorly understood. A major event, such as an insertions or a deletion of a segment of the amino acid sequence, may move an evolving protein to a substantially new area of its fitness neighborhood. Here, we studied the amino acid evolution caused by indels of one or more amino acids in proteins of Drosophila. An average insertion triggers ~0.5 amino acid replacements in the 10-20 amino acids surrounding the site of insertion. An average deletion affects a wider region (100-150 amino acids) and triggers ~3.5 amino acid replacements. The amino acid replacements triggered by insertion (deletion) usually - in 80% (65%) of cases - occurred 5' of the indel. Compared to the insertions, deletions had a lower ratio of fixed alleles to polymorphic alleles, and allele frequency spectrum skewed more to the left, suggesting stronger negative selection against deletions. Therefore, deletions appear to be more deleterious than insertions, and many of the adaptive amino substitutions that surround them may compensate for their deleterious effect. In summary, our results show that radical changes in amino acid sequence may be followed by a trail of adaptive substitutions.

Iterative *in-silico* and *in-vitro* tools for exploration of bitterness

Anat Levit[1], Ayana Wiener[1], Claudia Deutschmann[2], Maik Behrens[2], Wolfgang Meyerhof[2] and Masha Y. Niv[1]

[1] *Institute for Biochemistry, Food Science and Nutrition, The Robert H. Smith Faculty of Agriculture, Food and Environment, The Hebrew University of Jerusalem, Rehovot, Israel.* [2] *Department of Molecular Genetics, German Institute of Human Nutrition, Potsdam-Rehbruecke, Germany*
niv@agri.huji.ac.il, anat.levit@mail.huji.ac.il

Bitter-taste perception in humans is mediated by 25 GPCRs of the hTAS2R gene family. How can merely 25 receptors detect thousands of structurally diverse bitter compounds, and why are some of the receptors broadly-tuned, while others are capable of binding only a small number of ligands? The structural basis for the bitter taste receptors unique ability to specifically allocate numerous chemically diverse agonists is the focus of the current study.

To elucidate the sites of interaction between the bitter taste receptors and its different agonists, we generate an all-atom 3D model of the receptor. Comparative sequence conservation analysis for each of the binding site positions in the hTAS2R family is then performed to determine which residues may contribute to ligand binding and specificity. Computational docking of this ligand to the receptor is used to evaluate feasibility of the predicted specific interactions and is followed by *in-vitro* assays to confirm the proposed binding mode. This approach was applied to the broadly-tuned hTAS2R14 bitter taste receptor. Functional assays on wild-type vs. mutant constructs confirmed the *in-silico* predicted interactions and corroborated the main predicted binding site, situated inside the trans-membrane bundle. This site is analogous to previously identified binding pockets of other bitter taste receptors, such as the PTC bitter taste receptor (hTAS2R38) which was determined by means of homology modeling and hTAS2R46 which was elucidated by a combination of experimental and computational means, and also to non-bitter GPCRs for which structures were determined experimentally (Rhodopsin, β1 and β2-adrenergic receptors and A_{2A}-adenosine receptor).

The information on bitter compounds and receptors from the literature and from on-going experiments is organized in our database BitterDB. Since there is no free public database which specializes in taste compounds, the aim of BitterDB is to (1) gather all available public data on bitter molecules and their corresponding receptors; (2) characterize bitter molecules and computationally predict additional bitter candidates; (3) predict receptor functional groups involved in ligand binding. Currently BitterDB holds over 300 chemical structures of bitter tastants, and is constantly growing.

The iterative scheme for elucidation of molecular recognition of the bitter compounds by their cognate receptors represents first steps towards in-silico bitterness prediction.

Cytoplasmic Male Sterility: Can Microarray Help Us?

Alexei Levitchi, Rodica Martea, Daniela Abdusa, Maria Duca

University Center of Molecular Biology, University of the Academy of Sciences of Moldova, Moldova, levitsky_alex@yahoo.com

University Center of Molecular Biology, University of the Academy of Sciences of Moldova, 3/2 Academiei str., MD2028, Chisinau, Republic of Moldova, lab.bi.unasm@gmail.com

Cytoplasmic Male Sterility (CMS) represents underdevelopment of the pollen in flowers, which makes them just "females". It is determined by recombination events in mitochondrial genome. This effect is used widely in plant breeding for directed pollination between valuables genotypes, for hybrid production. Besides natural occurrence of CMS, was shown the possibility of its induction, by gibberelline (GA) treatment. There are no much evidences regarding neither natural, nor induced CMS mechanism. Thus, it was decided to reveal genes which are susceptible under GA treatment and their potential implication in the mechanism.

Gene expression database of microarray datasets served as data source (GSE8739). It was important to elaborate a pipeline of their exploratory analysis, testing the possibility of hypothesis elaboration regarding gene expression susceptibility to GA treatment and involvement in induced ASC (fig.1).

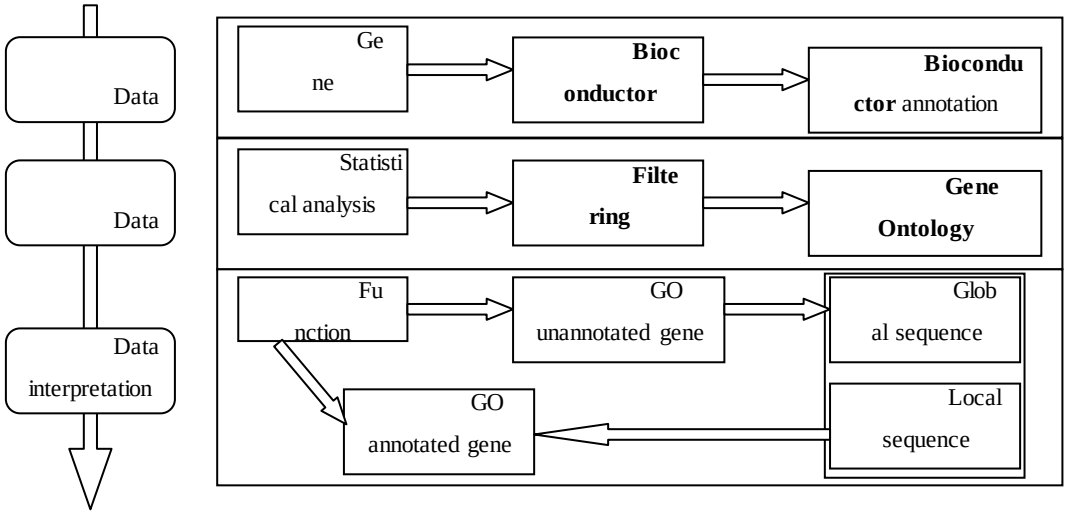


Figure 1. Workflow pipeline proposed for the study

Elaborated pipeline represents basic modules of activities: data extraction, their analysis and interpretation. Used tools represent R language based packages stored on Bioconductor.

From over 24000 genes presented on the chip, only 1600 were filtered according to expression fold change. Functional annotations from Gene Ontology revealed only 681 annotated genes. This gene set was grouped according GO terms, in order to identify their organelle distribution. All genes formed three groups, the biggest being *Cell part* (88% of genes). It comprises four groups, 45% of the genes representing *Intracellular* category. It is composed by 13 terms referred to different cell compartments. The majority of genes belong to *Nucleus*, *Plastid* and *Mitochondrion*. There were assumed three groups: Nucleus, Mitochondrion and Cytoplasm (fig. 2). The last is the environment, which contain all other terms. It is still a need to validate the genes referred to each of the category, through searching KEGG or AraCyc databases.

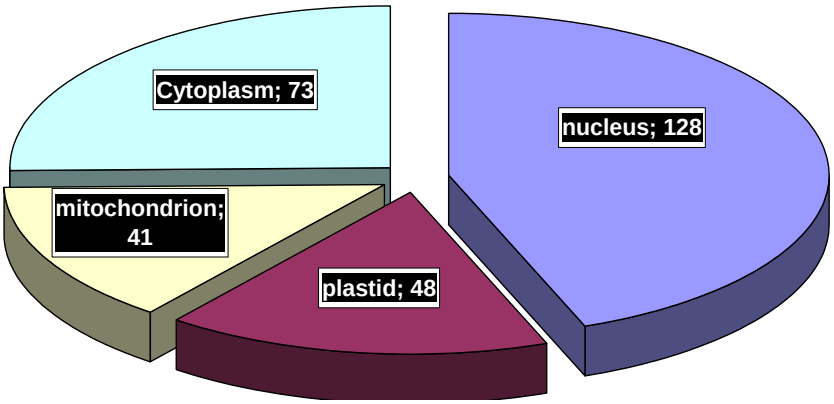


Figure 2. Number of genes referred to each of the

Another problem to be solved is unannotated sequences. As, it is not possible to estimate directly their function, it is proposed to verify two possibilities by alignment. Global alignment will be done using BLAST tool, while local will be done for motif and domain search. Based on these results, it might be possible to attribute functions to unannotated sequences. It will help enrichment of the primer dataset.

The research is done in frame of the National Institutional Project "Functional and genetic molecular aspects of sunflower (Helianthus annuus L.) genome"

1 T. Barrett (2011) NCBI GEO: archive for functional genomics data sets—10 years on, Nucleic Acids Research, **39**:1005-1010
2. M. Ashburner, C. A. Ball, J. A. Blake et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium, Nature Genetics, **25**(1):25-29
3.R.Gentleman et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biology, **5**:R80

Reference collection of transcriptional regulons in Bacillales family of bacteria

Semen A. Leyn[1,2], Marat D. Kazanov[2], Pavel S. Novichkov [3], Dmitry A. Rodionov[1,2]

[1]*Sanford-Burnham Medical Research Institute, La Jolla, California 92037;*
[2]*Institute for Information Transmission Problems RAS (Kharkevich Institute), Moscow 127994, Russia;*
[3] *Lawrence Berkeley National Laboratory, Berkeley, California, 94720*
semen.leyn@iitp.ru; rodionov@burnham.org

Gram-positive facultative anaerobic bacteria from the *Bacillales* order were isolated from diverse habitats including soil, sea water, plants and animals. *Bacillales* use diverse strategies to respond and survive in a variety of stresses and environmental conditions including resistance to multiple antibiotics. *Bacillus subtilis* subsp. subtilis str. 168 is one of the best-characterized Gram-positive bacteria and a model organism for studying sporulation, cell differentiation, stress response and social behavior of bacteria. According to the DBD database, *B. subtilis* genome encodes 238 DNA-binding transcription factors (TFs) classified in 45 protein families. Of them, 120 TFs were studied experimentally and the respective regulatory interactions were captured in the DBTBS database [1]. However, many of these known TF regulons have been studied insufficiently, providing an incomplete knowledge of TF-regulated genes and/or often unknown TF-binding sites (TFBSs).

In this study, we utilized the comparative genomics approach, as implemented in the RegPredict web-server [2], to reconstruct the transcriptional regulation network in the genome of *B. subtilis* and 10 related genomes from the *Bacillales* order. As result, we inferred a reference collection of 121 regulons for DNA-binding TFs and 32 regulons that operate using conserved RNA regulatory elements. As an input for the regulon reconstruction procedure we used any experimental information on transcriptional regulation in *B. subtilis* collected from more than 300 papers and the DBTBS database.

For analysis of TF regulons, first we re-analyzed and expanded 57 TF regulons with previously known TFBS sites in *B. subtilis* and propagated them to all studied genomes, resulting in refinement of TFBS motifs and identification of novel regulon members. Second, we identified novel TFBS motifs and described regulons for 28 experimentally studied in *B. subtilis* regulators with previously unknown binding sites. Thirdly, we discovered novel TFBS motifs and reconstructed regulons for 36 previously uncharacterized TFs. These novel regulons predicted to control genes involved in the following biological processes: utilization of various carbohydrates (alpha-galactoside, beta-glucoside, sucrose, inositol, maltodextrin, rhamnose and

rhamnogalacturonan); metabolism of glutamate, histidine, and thiamine; stress responses; drug/metabolite transport. Most of the identified TFBS motifs have either palindromic or tandem repeat structure suggesting that the respective TFs bind DNA as dimers or oligomers. Totally, more than 3000 TFBSs have been predicted in the *Bacillales* group (from 150 to 400 sites per genome).

For analysis of RNA regulons, we used bacterial RNA regulatory motifs collected in the Rfam database and scanned the studied genomes to identify new occurrences of RNA elements of each type. Then, the genomic context of the identified RNA elements was analyzed using the RegPredict approach, resulting in reconstruction of 32 RNA regulons. The reconstructed RNA regulons are operated by 11 families of metabolite-sensing riboswitches, 15 types of aminoacyl-tRNA-responsive T-boxes, one regulon controlled by the RNA-binding protein PyrR, and 5 regulons for experimentally uncharacterized RNA motifs. The RNA motif regulons in Bacillales control biosynthesis of vitamins and cofactors (cobalamin, riboflavin, thiamine, nucleoside queuosine), biosynthesis of glucosamine, metabolism of most amino acids, biosynthesis and salvage of purines and pyrimidines, and magnesium homeostasis.

The reference collection of transcriptional regulons for the Bacillales group of bacteria is available in the RegPrecise database (http://regprecise.lbl.gov).

1. N. Sierro et al. (2008) DBTBS: a database of transcriptional regulation in Bacillus subtilis containing upstream intergenic conservation information. *Nucleic Acids Res*., 36:D93-D96.
2. P. Novichkov et al. (2010) RegPredict: an integrated system for regulon inference in prokaryotes by comparative genomics approach. *Nucl. Acids Res*., 38: W299–W307.
3. P. Novichkov et al. (2010) RegPrecise: a database of curated genomic inferences of transcriptional regulatory interactions in prokaryotes. *Nucl. Acids Res*., 38: D111-8.

Comparative genomic reconstruction of N-acetylgalactosamine catabolic pathways and transcriptional regulons in Proteobacteria

Semen Leyn[1,2], Fang Gao[3], Chen Yang[3], Dmitry Rodionov[1,2]

[1]*Sanford-Burnham Medical Research Institute, La Jolla, California 92037;*
[2]*Institute for Information Transmission Problems RAS (Kharkevich Institute), Moscow 127994, Russia;*
[3]*Key Laboratory of Synthetic Biology, Institute of Plant Physiology and Ecology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200032, China*
semen.leyn@iitp.ru, rodionov@burnham.org

The bioinformatics approach was applied to reconstruct the N-acetylgalactosamine (GalNAc) and galactosamine (GalN) utilization pathways and the cognate AgaR regulon in genomes of Proteobacteria. The AgaR repressor has been previously characterized in *Escherichia coli* [1]. It belongs to the DeoR family of regulators and negatively controls the expression of the *aga* gene cluster in response to GalNAc and GalN in the medium. AgaR binds in tandem to several repeat sequences in the intergenic regions of *agaZ*, *agaR*, and *agaS* to repress transcription by overlapping the -35 and -10 boxes.

We identified orthologs of AgaR protein and reconstructed the respective regulons in 6 taxonomic groups of Gamma-proteobacteria (Enterobacteriales, Vibrionales, Pasteurellales, Alteromonadales, Aeromonadales, Xanthomonadales), and in two species that belong to Beta-, and Alpha-proteobacteria (*Burkholderia cenocepacia* J2315, *Caulobacter* sp. K31). In all cases, the *agaR* gene was located inside or in close proximity with the *aga* gene clusters encoding the components of GalNAc/GalN catabolic pathway. The paraphyletic structure of the maximum likelihood phylogenetic tree of AgaR proteins suggests the involvement of multiple horizontal gene transfer events in the evolution to *agaR*/*aga* loci. This hypothesis is also confirmed by the phylogenetic analysis of the GalNAc utilization enzymes and by the local character of the reconstructed AgaR regulons, when all predicted AgaR-binding sites are located in the proximity to the *agaR*/*aga* loci.

Using comparative genomic analysis we identified 4 distinct types of candidate AgaR-binding DNA motifs associated with the *aga* genes. All four motifs are repeats that share a common sequence pattern, CTTTC. For two groups of AgaR orthologs, their cognate DNA motifs can be described as direct repeats with conserved 10-nt distance between them. In contrast, the *Shewanellaceae* group has the predicted AgaR-binding motif that was defined as a 19-nt palindrome (nAAACTTTWWWAAAGTTTn), which is partially similar to the above

direct repeat motif. For the fourth group of AgaR orthologs, the regulatory regions of the *aga* operons contain several conserved regions with multiple CTTTC motifs found either in back-to-back or back-to-front orientations with flexible length spacer between them.

The predicted AgaR-regulated genes have been functionally annotated, resulting in reconstruction of the GalNAc and GalN utilization pathways in the studied microorganisms. The most conserved members of AgaR regulons are *agaZ* (tagatose-6P kinase) and *agaS* (predicted GalN-6P isomerase), whereas *agaI* (previously annotated as a GalN-6P isomerase), was found only in two studied species, *E. coli* and *Enterobacter*, suggesting that it has an auxiliary function. The combined phylogenetic and genome context analysis of AgaR-regulated sugar uptake phosphotransferase systems (PTS) was used to assign either GalNAc or GalN as their cognate substrates. In addition, the GalN-disaccharide specificity was assigned to a single PTS in *Haemophilus parasuis*, which has an unusual composition of the *aga* locus that was likely acquired horizontally from the *Streptococcaceae* species. In the genomes with missing GalNAc PTS systems, we predicted novel GalNAc permeases and kinases of various types that are all regulated by the AgaR regulons. In *Shewanella* species, genes encoding predicted GalNAc-specific transporter *agaP* and deacetylase *agaA2* were identified as paralogs of the respective *nagP* and *nagA* genes encoding the N-acetylglucosamine-specific transporter and deacetylase, respectively, suggesting likely involvement of recent duplications and changes in specificities in the evolution of these proteins.

The novel variant of GalNAc pathway identified in *Shewanella oneidensis* MR-1 has been experimentally validated *in vitro* using enzymatic assays with the purified recombinant proteins AgaS, AgaK and AgaA2.

1. Ray, W. K. and Larson, T. J. (2004), Application of AgaR repressor and dominant repressor variants for verification of a gene cluster involved in N-acetylgalactosamine metabolism in Escherichia coli K-12. *Molecular Microbiology*, 51: 813–826

GPCR – Ligand Docking with Refining Receptor Interface

Lingyun YANG[1], Zhonglan LUAN[1], Qiang LU[1,2], Xiaoyan XIA[1,2]

[1] *School of Computer Science and Technology, Soochow University, Suzhou, 215006, China , qiang@suda.edu.cn*
[2] *Jiangsu Provincial Key Lab for Information Processing Technologies, Suzhou, 215006, China*

G-protein coupled receptors (GPCRs) play an important role in signal transduction and are targets for almost half of the existing drugs. However, 3D homology modeling is still the major source of GPCRs structural models for rational drug design because it is hard to solve GPCR crystal structure. So when computational docking a ligand into a GPCR, the modeling of conformational flexibility of receptor can help to predict accurate GPCR-ligand binding mode.

The paper presents a method with simultaneous refinement of receptor's interface and docking process. The receptor's interface is defined as the residues whose at least CB (CA for GLY) atom is with ≤ 7Å distance to any heavy atom of the ligand. We modified RosettaLigand[1,2] docking algorithm to incorporate our refining procedure. We set our refinement procedure before the second stage or the high-resolution refining stage[2] of the original RosettaLigand algorithm. After identifying the receptor's interface, we apply kinematic closure (KIC) algorithm [4] to refine those loops' at interface. Figure 1 shows that our refinement procedure is embedded before the High-resolution refining stage.

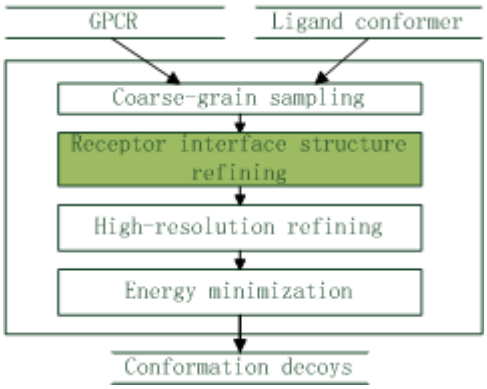


Figure 1 The layout where the refinement of receptor interface resides into RosettaLigand

We use RosettaLigand (RL) and our method (RLEN) to redock five A2A-ZMA (Human A2A adenosine receptor with ligand ZM24125, PDB ID: 3EML), starting with the computiaonal models from Baker Group on GPCR Dock 2008[3]. Five A2A-ZMA modes are named from A2A_1 to A2A_5. CA RMSD and LRMSD (in angstrom) of the complexes are: A2A_1(3.5, 5.5), A2A_2(4.6, 7.1), A2A_3(4.2, 7.1), A2A_4(3.9, 9.9), A2A_5(3.9, 12.2).For each starting

model, we use RLEN and RL to generate 1000 decoys and compare the 10% lowest LRMSD decoys. We found that for low-accuracy starting models, RLEN can generate more decoys with lower LRMSD. For example, Figure 2 shows the results for redocking A2A_4 and A2A_5 with ZMA. We can see that RLEN performs better than RL especially for A2A_4 that the docoys generated by RLEN with LRMSD between 3.0 and 3.5 are more than twice of by RL.

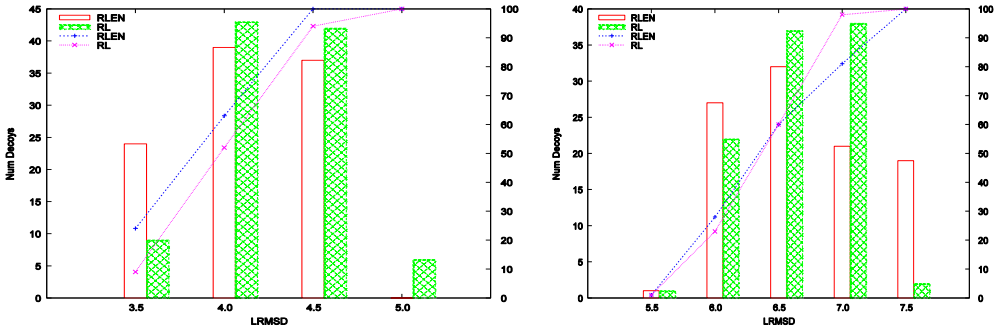


Figure 2 Results of A2A_4 (left) and A2A_5 (right)

1. J.Meiler et al. (2006) RosettaLigand: Protein-small molecular docking with full side-chain flexibility, *PROTEINS*, **65:**538–548.
2. I.W.Davis et al. (2009) RosettaLigand docking with full ligand and receptor flexibility, *J. Mol. Biol.*, **8:**455–463.
3. M.Michino et al. (2009) Community-wide assessment of GPCR structure modeling and ligand docking: GPCR Dock 2008, *Nat. Rev. Drug. Discov.*, **8:**455–463.
4. D.J.Mandell et al. (2009) Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformation sampling, *Nature Methods*, **6**(8)**:**551-552.

Tandem repeat polymorphisms in the human genome

Dmitrijs Lvovs[1], Vsevolod Makeev[2,1], Marina Fridman[2], Nina Oparina[3]

Research Institute for Genetics and Selection of Industrial Microorganisms, Moscow, Russia, dmitrijs.lvovs@gmail.com, makeev@genetika.ru, marina-free@mail.ru
[2] Vavilov Institute of General Genetics, Russian Academy of Sciences, Moscow, Russia
[3] Engelhardt Institute of Molecular Biology, Rusisan Academy of Sciences, Moscow, Russia, oparina@gmail.com

Tandem repeat (TR) polymorphisms provide one of principle sources of genomic variability. They have been reported to modulate a range of biological processes (Hannan, 2010). The spectrum of polymorphisms is determined by biases between different mutation types and subsequent natural selection. Thus, the study of these polymorphisms is an efficient tool for studying variability associated with TRs (Fu and Chakraborty, 1998) and functions of these repeats (Hannan, 2010).

We examined occurrence of different types polymorphisms (insertions and deletions, SNPs, microsatellite and others) in TRs presented in the UCSC database: ([simplerepeat] table for TRs and [snp130] table for other polymorphisms). We compared [SNP130] records with TRs (starting and ending positions of the polymorphism reference were taken at each chromosome and compared with TR positioning), and performed exhaustive search for hypotheses of association between different parameters of [SNP130] and TR records. Fisher exact test and Odds Ratio were used for inference. The ‘class’ and ‘length’ fields in [SNP130] table and the ‘period’ of TRs were used for association parameters. To reduce the size of the resulting data set, we filtered records keeping only those with Period and Length less than 24 b.p. We obtained a series of strong Fisher p-values < E-200, which allowed us to use Bonferroni correction for multiple testing.

Many strong associations (Fisher p < E-200) were found for SNP lengths multiple to TR period, a phenomenon which is expectable but was never demonstrated directly at the genome level.

This agrees with the view that in TRs mutations occur in a stepwise manner (insertion/deletion of one, sometimes two units of repeat). It is noteworthy that this is true not only for microsatellites but also for minisatellites. The mechanisms of mutation of these TR types are different one from another (polymerase slippage vs recombination) (Y.-X.Fu and R.Chakraborty,1998, G.Vergnaud and F.Denoeud, 2000), but typical SNP’s are similar.

Also, some interesting trends have emerged after cross-sectional data analysis, that have to be inspected in more detail. For example, the most frequent repeat periods that have appeared in results were period 11 and 23, all such associations with Odds Ratios less than 1, proposing some deeper investigation of those TR periods. This tendency, however disappeared when we

considered only the exact matches of TRs, as proposed by TRF algorithm, which may indicate some biases in TRF algorithm.

Our study demonstrates strong association of polymorphisms in human genome with frequent tandem repeats. By the way, due to experimental procedures current dbSNP includes mostly data on single nucleotide variations, not for such polymorphisms like microsatellite copy number variations. That is why further studies of human genome resequencing projects could give additional information on frequency of different SNP types in human population.

This work was partially supported by the EC Seventh Framework Programme (FP7/2007-2013), grant agreement #212877 (UEPHA*MS)

1. Y.-X.Fu and R.Chakraborty (1998) Simultaneous Estimation of All the Parameters of a Stepwise Mutation Model, Genetics 150: 487–497.
2. A.J. Hannan (2010) TRPing up the genome: tandem repeat polymorphisms as dynamic sources of genetic variability in health and disease, Discov Med , 10(53):314-321.
3. G.Vergnaud and F.Denoeud (2000) Minisatellites: Mutability and Genome Architecture, Genome Res., 10: 899-907.

Bacterial type RNA polymerase sigma subunits and their specific promoters in plastids

K.V. LOPATOVSKAYA, A.V. SELIVERSTOV, V.A. LYUBETSKY

IITP RAS, Russia, 127994, 19, B. Karetny per., Moscow,lyubetsk@iitp.ru

There are three phyletic lineages with plastids: the Red Line, Green Line and *Cyanophora paradoxa*. *Arabidopsis thaliana* is known to possess six types of sigma subunits, while in distant taxa minor sigma subunits are poorly studied. In this research, data was obtained from NCBI GenBank and the Sanger institute. An original method and its implementation are developed to search for promoters based on estimating the transcription initiation frequency at an arbitrary DNA motif. In particular, the results below are interesting.

The Red Line. In diatom algae four sigma subunits are found: ThSig1a, ThSig1b, ThSig2, and ThSig3, with ThSig1a and ThSig1b being closes paralogs. The complete set occurs only in *Thalassiosira pseudonana*. *Phaeodactulum tricornutum* lacks the ThSig1a paralog. Close homologs of ThSig1a,b are found in the brown alga *Ectocarpus siliculosus* and the raphidophyte *Heterosigma akashiwo*. In cryptophytes and rhodophytes the ThSig1a,b orthologs are not found. In cryptophytes ThSig2 has one ortholog in *Hemiselmis andersenii* and two close orthologs in *Guillardia theta*. ThSig3 has two orthologs in rhodophytes. Many close subunit orthologs are found in Rhodophyta, e.g., among the five in *Cyanidium caldarium*, SigB and SigC differ only in

11 amino acid substitutions. *E.siliculosus* possesses additional two distantly related subunits. Several sigma subunits exist in the pelagophyte alga *Aureococcus anophagefferens*. Apicomplexan parasites *Toxoplasma gondii*, *Eimeria tenella*, *Plasmodium falciparum* 3D7, *Pl. yoelii str.* 17XNL, *Pl. vivax* SaI-1, *Pl. Chabaudi*, *Pl. knowlesi*, *Pl. berghei* are found to possess a single sigma subunit type.

Glaucocystophyceae. The SigA and SigB subunits in *C. paradoxa* are significantly divergent, particularly at the N-end, although containing conserved regions as well; SigA is close to ThSig2 in diatoms, SigB has no evident orthologs in algae.

Early lineages of the Green Line. *Bigelowiella natans* and most other Chlorophyta are found to possess a single sigma subunit, with the exception of *Micromonas pusilla* and *Chlorella variabilis*, which both have two closely related subunits. Little subunit divergence within each species suggests their recent independent origins through duplications.

Streptophyta. The Sig1 subunit is typically present in all photosynthetic species, while *Nicotiana tabacum*, *Populus trichocarpa*, *Vitis vinifera* have two close paralogs, Sig1A and Sig1B. Two close paralogs of Sig2, Sig2A and Sig2B, are found in *P. trichocarpa* and the Poales *Oryza sativa, Sorghum bicolor, Zea mays;* a single Sig2 subunit exists in most rosids, in *Spinacia oleracea* and in early land plants, e.g. conifers, lycopodiophytes and mosses. Distant homologs of Sig2 are found in *Solanum lycopersicum* and *Artemisia annua* but lack in *N. tabacum*. Sig4 occurs in many rosids species with Sig4-dependent promoters. In land plants Sig5 tends to co-occur with the plastid transcription factor 1. The Sig6 protein is restricted to Poales and rosids, with its distant homologs found in *S. lycopersicum* and *Picea glauca*.

Plastid promoters. Only five highly conserved promoters are found in Streptophyta. In diatoms and related plastids, promoters upstream of some genes, including *psaA*, possess two G bases in their -35 boxes that differ from both the bacterial and taxon specific promoter consensus sequences. The predicted subunits ThSig1a, ThSig1b, ThSig2 and ThSig3 might specifically bind to these promoters. However, bacterial type promoters with at least two consensus sequences were found in diatom algae.

This study focuses on the evolution of sigma subunits and their specific promoters in plastids of plants and algae. The subunits - promoters complex differs considerably between different algal groups and between algae and higher plants. In Apicomplexa, the bacterial and phage RNA polymerase transcript ratio was studied, with potential implications in predicting drugs resistance under selective suppression of the bacterial type polymerase. Detailed results will be presented.

Research was partly supported by the Ministry for Education and Science of Russia (grants P2370 and 14.740.11.0624).

BioinfoWF — PLATFORM FOR RAPID DESIGN OF THE WEB SERVICES AND WORKFLOWS FOR BIOINFORMATICS ANALYSIS

Genaev M, Gunbin K, Afonnikov D

Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia, Russian Federation, mag@bionet.nsc.ru

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

To perform workflow data processing for bioinformatics we developed BioinfoWF system (http://pipeline.bionet.nsc.ru/). The system consists of server and client sides. The server side is implemented in Perl. It starts workflow execution and monitors module calculation status. The workflow server allows performing resource-intensive tasks at the HPC cluster. The system also performs running modules in parallel in case workflow topology allows it. The input of the system is data and pipeline description in XML format. The workflow description consists of two files, a list of modules and topology description.

The client part of the application is implemented as a web-application. The user interface for a particular workflow is automatically generated by the system. User can navigate the workflow scheme, modify it and set input data for each module.

The BioinfoWF system was utilized for design and implementation of the SAMEM package (http://pixie.bionet.nsc.ru/samem/) of the molecular evolution analysis of genes and proteins [3]. The system consists of two workflows. The first workflow performs the following tasks: (1) Multiple alignment nucleotide sequences of genes; (2) Translation of the coding nucleotide sequences into amino acid sequences; (3) Multiple sequences alignment; (4) Phylogenetic tree construction; (5) Estimation of parameters of the evolutionary models for proteins; (6) Estimation of parameters of the evolutionary models of nucleotides sequences; (7) Ancestor sequence inference. The second workflow performs similar analysis for amino acid sequences. The tasks at each workflow module can be solved by a number of methods, depending on the user's choice. For example, multiple sequence alignment could be performed by either MAFFT [4] or Kalign[5] programs.

The work was supported by RFBR grant No. 09-04-01641-a and Biosphere Origin and Evolution program.

1. Ríos J., Karlsson J. and Trelles O. (2009) Magallanes: a web services discovery and automatic workflow composition tool, *BMC Bioinformatics*, **10**:334.
2. Oinn T., Addis M., Ferris J., Marvin D., Senger M., Greenwood M., Carver T., Glover K., Pocock M.R., Wipat A., Li P. (2004) Taverna: a tool for the composition and enactment of bioinformatics workflows, *Bioinformatics*, **20**:3045-3054.
3. Gunbin K.V., Genaev M.A., Afonnikov D.A., Kolchanov N.A. (2010) A Computerized System for the Analysis of Molecular Evolution Modes of Protein-Encoding Genes (SAMEM): the Relationship between Molecular Evolution and Phenotypic Traits, *Moscow University Biological Sciences Bulletin*, **4(65)**:143-145.
4. Katoh K., Toh H. (2008) Recent developments in the MAFFT multiple sequence alignment program. *Brief Bioinform*, **9(4)**:286-98.
5. Lassmann T., Sonnhammer EL. (2005) Kalign – an accurate and fast multiple sequence alignment algorithm, *BMC Bioinformatics*, **6**:298.

BioUML: the ChIPMunk Plugin for Motif Discovery in ChIP-Seq Data

Vsevolod J. MAKEEV[1*], Ivan V. KULAKOVSKIY [2], Ivan S. YEVSHIN [3], Tagir F. VALEEV [3,4]

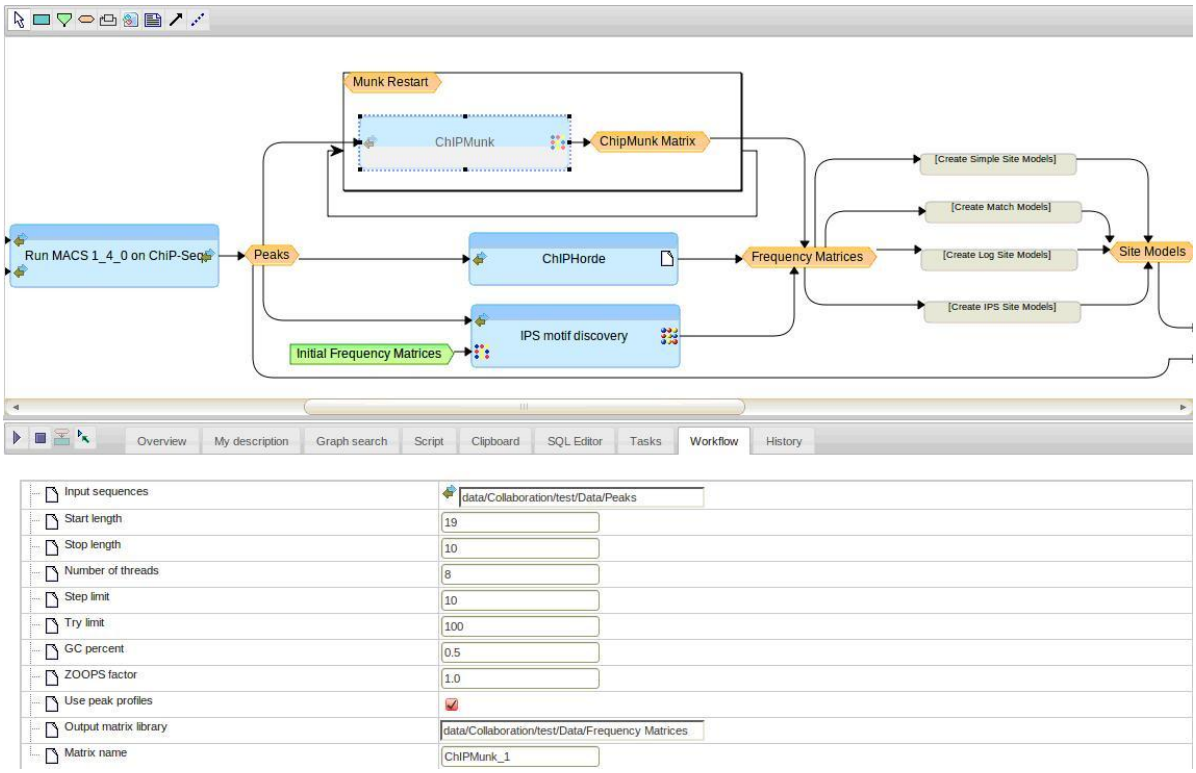
[1] *Vavilov Institute of General Genetics, RAS, Gubkina str. 3, Moscow 119991, Russia*
[2] *Engelhardt Institute of Molecular Biology, RAS, Vavilov str. 32, Moscow 119991, Russia*
[3] *Institute of Systems Biology, Detskiy pr. 15, Novosibirsk 630090, Russia*
[4] *Institute of Informatics Systems, SB RAS, Acad. Lavrentjev pr. 6, Novosibirsk 630090, Russia*
vsevolod.makeev@gmail.com
* *corresponding author*

Background. The task of identification of transcription factor binding motifs in a set of short DNA sequences has a long history. A new challenge for the motif discovery was recently provided by a huge amount of data from ChIP-Seq experiments. ChIP-Seq data can contain thousands of sequences where one needs to find a short overrepresented motif. Fortunately, ChIP-Seq data have additional information, the so-called 'base coverage profile', which represents the shape of a ChIP-Seq 'peak' and helps to select the correct signal.

Algorithm. We have constructed an efficient motif discovery tool called ChIPMunk. It combines greedy optimization with bootstrapping and allows one to make use of the base coverage profile information. Our own [1] and independent benchmarks [2] using several ChIP-Seq datasets have shown that ChIPMunk motif recognition quality is the same or better than that of the traditional (MEME, Multiple EM for Motif Elicitation) or ChIP-Seq oriented (HMS, Hybrid Motif Sampler) tools while the speed is dramatically better.

BioUML integration. BioUML platform [3] allows third-party developers to create their own plugins to create a new specific functionality in the system. We created a wrapper plugin to use ChIPMunk within BioUML. Such integration provides a number of benefits: 1) User-

friendly BioUML GUI (both standalone and web-based) instead of the command line interface; 2) Support of various data input formats, like sequence formats (Fasta, EMBL, Genbank), genome intervals (BED, GFF) or results of other analyses integrated in BioUML (e.g. MACS analysis [3]) along with peak profile information; 4) Use of resulting matrices to construct various binding site models to search for binding sites; 5) Motif logo visualization; 6) Research automation either using JavaScript or creating BioUML workflow in the visual editor. An example of workflow including ChIPMunk analysis is shown below:



This workflow allows to find peaks using MACS, feed them into various motif discovery methods including ChIPMunk and construct various site prediction models (PWM, logarithmic PWM, Match [5], IPS [6]). ChIPMunk parameters are shown in the lower part.

1. I.V. Kulakovskiy *et al.* (2010) Deep and wide digging for binding motifs in ChIP-Seq data, *Bioinformatics*, **26(20):**2622-3.
2. L. Kuttippurathu *et al.* (2011) CompleteMOTIFs: DNA motif discovery platform for transcription factor binding experiments, *Bioinformatics*, **27 (5):**715-717.
3. http://biouml.org/
4. Y. Zhang *et al.*(2008) Model-based analysis of ChIP-Seq (MACS), *Genome Biol*, **9(9):**R137.
5. A.E. Kel *et al.* (2003) MATCH: A tool for searching transcription factor binding sites in DNA sequences, *Nucleic Acids Res*, **31(13):**3576-9.
6. I.S. Yevshin *et al.* (2011) IPSscan: the extended matrix method for prediction of transcription factor binding sites, *Bioinformatics* (submitted).

Prediction of Genome-wide Interactions reveals Communication Signals during Mycobacterial Latency

Shubhada Hegde[1], Chandrani Das[2], Shekhar Mande[1]

[1]*Centre for DNA Fingerprinting and Diagnostics, Hyderabad, India, shubhada@cdfd.org.in*
[2]*Tata Consultancy Services, Hyderabad, India*

Mycobacterium tuberculosis is the causative agent of Tuberculosis which claims around 2 million deaths annually. About 90% of the people infected with Mycobacterium tuberculosis carry latent bacteria which are believed to get activated upon immune suppression [1]. One of the challenges is to understand the molecular mechanisms of latency and reactivation. A major obstacles in this regard is that in the completely sequenced genome of M. tuberculosis a large fraction of genes are either putative or are unannotated [2]. Thus, integration of different approaches and available individual as well as high-throughput datasets becomes inevitable to understand the biology of M. tuberculosis.

We have attempted to address the phenomenon of M. tuberculosis dormancy at systems level by analysis of genome-wide protein:protein interactions integrated with available large scale gene expression studies and the predicted transcription regulatory network. Based on genome-context methods, namely phylogenetic profile, gene distance and operonic frequency, and gene expression correlations, a prediction for genome-wide protein functional linkages was made using Support Vector Machine [3]. This set of protein functional linkages, along with gene expression data of the available models of latent M. tuberculosis, has been employed to identify proteins involved in mediating switch signals during dormancy. We identify 84 differentially regulated genes which are common among many gene expression studies of various dormancy models of M. tuberculosis. Interestingly, this set of genes that are up and downregulated during dormancy also exhibit inverse correlation in their expression across varied growth conditions indicating that they form a tightly regulated cluster. In addition, the down regulated genes are highly conserved in evolution whereas the upregulated genes are less conserved, suggesting a unique evolutionary history that might be associated with M. tuberculosis dormancy. The interaction profiles of M. tuberculosis proteins that are up and downregulated during latency were then examined to identify proteins that are involved in conveying dormancy signals between up and downregulated genes. Analysis of shortest paths between up and down-regulated genes in the network reveal proteins that might play important role in signal transmission during dormancy. The examples of such proteins include DosR and DosS, HspX, Rv2621c, SseC2 and the subunits of ATP synthase.

From our analysis we validate that DosR plays an important regulatory role in the dormancy switch and transmits dormancy signals to the respiratory system of M. tuberculosis. This eventually leads to the switching off of ATP sysnthesis and culminating in the significant slowdown of cellular growth and replication. We thus exemplify here the application of protein functional linkages and large scale data from sources such as microarray, in order to understand latency in M. tuberculosis. The list of predicted protein interactions and the associated datasets are available on our web server (http://www.cdfd.org.in/MtbPPI/).

1. Stewart GR, Robertson BD and Young DB. (2003) Tuberculosis: a problem with persistence. Nat Rev Microbiol. 1:97-105.
2. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, et al. (1998) Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. Nature. 393:537-544.
3. Yellaboina S, Goyal K and Mande SC. (2007) Inferring genome-wide functional linkages in E. coli by combining improved genome context methods: Comparison with high-throughput experimental data. Genome Res. 17: 527-535.

Towards understanding the gut microbiota of a malnourished child

Sharmila S Mande*, Monzoorul Haque Mohammed, Tarini Shankar Ghosh

Bio-Sciences R&D Division, TCS Innovation Labs Hyderabad, Tata Consultancy Services Limited, Hyderabad, India,
Email: sharmila@atc.tcs.com

G. Balakrish Nair, Sourav Sen Gupta, Suman Kanungo

National Institute of Cholera and Enteric Diseases, Kolkata, India,
Email: gbnair_2000@yahoo.com

Malnutrition is one of the major problems in developing countries. It affects around 13 million children world wide and accounts for 1-2 million deaths. The consequences of malnutrition are devastating and include diarrhoea, malabsorption, increased intestinal permeability and suboptimal immune response. The ineffectiveness of nutritional interventions and dietary solutions in treating this menace has prompted researchers to explore alternate avenues for addressing this issue. One of these avenues is to study the taxonomic and functional diversity of the microbial communities resident in the guts of malnourished children and compare the same with those associated with healthy children. Such studies are expected to not only help in understanding the microbial basis for several physiological disorders associated with

malnutrition (eg, increased susceptibility to diarrhoeal pathogens), but in the long run, can also help in devising appropriate pro-biotic strategies for eradicating this menace.

In this regard, metagenomics is rapidly becoming the method of choice for researchers as it facilitates the rapid extraction and characterization of the entire genomic content present in an environment, by bypassing the prior need for cultivation of microbes. Using metagenomic approaches, researchers can now access the complex cross-talk between the gut and its microbial flora and obtain a comprehensive understanding of how different community composition affects various states of human health. In this study, a metagenomic approach was employed for analysing the differences between gut microbial communities obtained from a malnourished and an apparently healthy child.

The taxonomic and functional diversity of metagenomes obtained from the guts of a malnourished and a healthy child were profiled and compared using several computational approaches. Our results revealed that the malnourished child gut has an abundance of enteric pathogens which are known to cause intestinal inflammation resulting in malabsorption of nutrients. The malnourished child gut was also observed to be selectively deficient of several bacterial groups like Lactobacillales and Bifidobacteriales, which are known to possess pro-biotic properties. In addition, a few pathogen-specific functional subsystems (or genes) were also identified, which probably impact the overall metabolic capabilities of the malnourished child gut.

The intestinal microflora of the malnourished child when compared to the healthy child is interpreted as aberrant gut microflora. Such an aberration leads to a subclinical disorder characterized by inflammation and modest malabsorption. Our study provides a preliminary picture of the gut microbial community resident in the gut of a child suffering from malnutrition. By providing a comprehensive snapshot of the microbial community resident in the gut of a malnourished child, this study has attempted to extend the understanding of the basis of malnutrition beyond nutrition deprivation.

Metagenomics Analysis Platform for Automatic Annotation of Metagenome Sequences obtained from Next Generation Sequencing Technologies

Monzoorul Haque Mohammed, Sudha Chadaram, CVSK Reddy, Sharmila Mande

Bio-Sciences R&D Division, TCS Innovation Labs Hyderabad, Tata Consultancy Services Limited, Hyderabad, India,

Keywords : metagenomics, analysis platform, algorithms, Next generation sequencing

The focus of Life Sciences and Health care R&D has been significantly impacted with the recent developments in the Next Generation Sequencing (NGS) technologies and the concomitant emergence of a new research area called 'Metagenomics'. The development in the new area of metagenomics has its roots in the following observation. Majority of microorganisms present in natural ecosystems cannot be cultured in laboratory. The metagenomics approach bypasses the culturing step and analyzes DNA obtained directly from microbes inhabiting a given environmental niche. This approach thus enables the direct exploration, characterization, and beneficial exploitation of the unexplored diversity of microorganisms. Results of several ongoing studies have indicated the tremendous potential of metagenomics in unearthing thousands of novel genes and proteins, some of which have the potential for commercial exploitation. NGS technologies have enabled rapid and simultaneous sequencing of millions/billions of DNA fragments in a cost effective manner. Given that genomic content of several thousand microbes (living in a particular environment) are sequenced and analyzed in a typical metagenomics project, NGS technologies have thus played a perfect complementary role in furthering the development of Metagenomics.

Despite the tremendous potential of NGS technologies and metagenomics, the lack of tools/algorithms/analysis platforms that can perform an accurate meaningful end-to-end analysis of generated data remains a prime concern of researchers in life sciences and health care sectors. Besides specialized algorithms, significant compute power and infrastructure are also required for analyzing metagenomics data.

We present a comprehensive metagenomics analysis platform developed at Tata Consultancy Services' (TCS) Innovation Labs, Hyderabad, India. TCS' metagenomic analysis platform implements a suite of in-house developed, published and patented algorithms. The platform includes algorithms for data pre-processing, decontamination of host associated sequences, detection and taxonomic characterization of 16S rDNA sequences, functional and taxonomic characterization of the entire metagenomic content, comparative analyses of metagenomic data sets as well as tools for the detection of habitat specific genes/sequences. The platform provides work-flow creation and customization capabilities which enables an end to end analysis of metagenomic data sets. Details of this metagenomics analysis platform will be presented during the conference.

Identification of partial MHC class II B exon 2 sequences in 3 European Ranidae species

Bela Albert Marosi[1], Ioan Valeriu Ghira[2], Tibor Sos[3], Octavian Popescu[1]

[1]*Molecular Biology Center, Interdisciplinary Research Institute on Bio-Nano-Sciences, Babes-Bolyai-University Cluj-Napoca, Romania, marosib@yahoo.com*
[2]*Faculty of Biology and Geology, Babes-Bolyai-University Cluj-Napoca, Romania*
[3]*Association for Bird and Nature Protection "Milvus Group, Romania*

The major histocompatibility complex (MHC) genes are the most polymorphic genes of the vertebrate genome and play a key role in the adaptive immune system. Due to their high variability they can be used as molecular markers for assessment of population genetic structure and populations' phylogeography. The use of MHC genes for species phylogeny is still disputed. In this study we report the partial MHC class II B exon 2 sequences of three Ranidae species: Rana arvalis, Pelophylax kurtmuelleri and Pelophylax lessonae, which according to our knowledge have not been described so far. For our study we used 3 individuals of Rana arvalis, 2 individuals of Pelophylax lessonae and 2 individuals of Pelophylax kurtmuelleri. Species were identified by morphological traits and in case of the two Pelophylax species we ensured identification with the amplification of a partial 16 S gene sequence from the 4 individuals. The 16S sequences of these frog species can be found in the NCBI database, thus the BLAST of our sequences proved that the morphological identification was correct. In case of all 3 species the degenerate MHC primer products were cloned and 10 colonies were analyzed for each individual. We do not claim that with this method we were able to find all the alleles and genes, but our initial aim was only the identification of the target sequences. For the Rana arvalis the target sequence length excluding primers was 186 bps and we found 8 different sequences. For the Pelophylax kurtmuelleri the target sequence length excluding primers was 196 bps and we found only one allele. For the Pelophylax lessonae the target sequence length excluding primers was 196 bps and we found 4 different sequences. Codons involved in antigen binding were identified as well.

A phylogenetic analysis of our sequences together with sequences of other frog species form the NCBI database indicates that there are overlapping alleles among species and in this case the MHC class II B exon 2 sequences are not a precise tool for species delimitation. Our findings open the possibility for further population analyses of these frog species, based on the MHC class II B gene sequences.

Identification of shortened 3'untranslated regions and impact on microRNA regulation

Loredana Martignetti[1,2,3], Karine Laud-Duval[1,4], Franck Tirode[1,4], Emmanuel Barillot[1,2,3], Olivier Dellatre[1,4] and Andrei Zinovyev[1,2,3]

[1] *Institut Curie, 26 rue d'Ulm, Paris, F-75248*
{loredana.martignetti, karine.laud, franck.tirode, emmanuel.barillot, olivier.delattre, andrei.zinovyev} @curie.fr
[2] *INSERM U900, Paris, F-75248*
[3] *Mines Paris Tech, Fontainebleau, F-77300*
[4] *INSERM U830, Paris, F-75248*

In eukaryotes, genes are regulated at many different levels to produce the correct assortment of proteins for every cell type. The discovery of RNA silencing pathways focused the attention on post-transcriptional control as a key layer of regulation in several biological processes. Untranslated regions of eukaryotic mRNAs contain motifs that are essential to regulate post-transcriptional processes (e.g. mRNA processing, export, surveillance, silencing by microRNAs and turnover). At the end of every mRNA there is a signal that indicates that the end of the mRNA is reached (the polyadenylation signal). In many genes, two or more polyadenylation signals are found in the 3′ UTR, so that different isoforms with different 3'UTR length can be expressed. This mechanism, called alternative polyadenylation (APA) is quite common in human mRNAs and it is subject to tissue or condition specificity [1]. Recently it has been shown that cancer cells often expressed substantial amounts of mRNA isoforms with shorter 3' UTRs [2]. This is relevant from the point of view of post-transcriptional regulation because if the 3'UTR of a mRNA is shorter or missing, miRNAs and other regulatory proteins are not longer able to bind.

We present here a computational procedure for systematically identifying APA events by Affymetrix GeneChip microarrays. The advantage of this technology compared with more recent and promising ones such as exon arrays and RNA-Seq is that, giving the relatively small cost, a typical study includes a considerably higher number of experiments. Moreover, the design of Affymetrix Gene Chips is well-suited for 3'UTR analysis of a large number of genes.

The proposed approach requires as input the expression profile from Affymetrix GeneChip array for the samples of interest. As final result of our analysis we obtain a set of genes expressing short 3'UTR isoform in a minimum number of the analyzed samples.

Initially, Affymetrix GeneChip single probes are assigned to CDS or 3'UTR of the transcript, according to NCBI RefSeq database annotation (Release 45). Then we define for each RefSeq two distinct meta probe sets, the first one including probes covering specifically the CDS

and a second one including probes covering specifically the 3'UTR. The expression ratio between these two meta probesets is expected to be equal to one in case the 3'UTR is not subject to shortening. A high value of CDS:3'UTR expression ratio is indicative of variation in the expression between the CDS and the 3'UTR and it can be interpreted as an event of short 3'UTR isoform expression.

The procedure has been applied to expression data from 75 samples of Ewing's sarcoma patients generated by Affymetrix U133A microarray. We used all sequences supported by RefSeq and required at least 4 probes in both CDS and 3'UTR meta probe sets for each gene. Among the 5500 genes selected in this way, we extracted a list of 266 genes showing short 3'UTR expression in at least 10% of Ewing's sarcoma patient samples. We checked whether the extracted genes have multiple annotated 3'UTR isoforms. The extracted gene list has been crossed with gene entries with multiple polyadenylation signals confirmed by both AltTrans and AltPas polyA site databases (2856 entries) [3]. The overlapping list contains 74 genes ($Pv \sim 10^{-9}$), confirming that our procedure enables us to identify candidate 3'UTR shortening events.

Further analysis are in progress to recognize potentially relevant cases in which truncated 3'UTR is responsible for loss of microRNA binding sites.

1. F Ozsolak, P Kapranov, S Foissac, SW Kim, E Fishilevich, AP Monaghan, B John, PM Milos. (2010) Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation. *Cell*. Dec 10;**143(6):**1018-29.
2. C Mayr, DP Bartel, (2009) Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell*. Aug 21;**138(4):**673-84.
3. V Le Texier, JJ Riethoven, V Kumanduri, C Gopalakrishnan, F Lopez, D Gautheret, TA Thanaraj. (2006) AltTrans: transcript pattern variants annotated for both alternative splicing and alternative polyadenylation. *BMC Bioinformatics*. Mar 23;**7:**169.

Effect of Bacillus thuringiensis Cry3Aa toxin on Tenebrio molitor transcriptome composition

Alexander Martynov[1], Darya Evsyutina[1], Brenda Oppert[2], Elena Elpidina[3]

[1]*Faculty of Bioengineering and Bioinformatics, Moscow State University, Russian Federation, agmart@mail.ru*
[2]*USDA ARS Center for Grain and Animal Health Research, United States*
[3]*A.N. Belozersky Institute of Physico-Chemical Biology, Moscow State University, Russian Federation*
elp@belozersky.msu.ru

Agriculture all over the world suffers great losses from insect pests. They damage crops and stored grains in storehouses. Chemical insecticides affect environment and often are not safe for human health. Biopesticides based on *Bacillus thuringiensis* (*Bt*) delta-endotoxins are considered to be among the most progressive and perspective. These insecticides are toxic for many insect pests from the orders Lepidoptera, Diptera, and Coleoptera, but are safe for humans and all Vertebrata. *Bt* toxins are widely used in USA, Spain and other countries, as well as transgenic plants expressing Cry-toxins. Coleopteran-specific Cry3A toxin is used for Colorado potato beetle control.

Yellow mealworm, *Tenebrio molitor*, is a pest of stored grains and grain products. It is susceptible to the *Bt* var. *tenebrionis* Cry3Aa toxin, although the toxic effect is not very strong. Our goal was to monitor the most remarkable changes in the transcriptome of *T. molitor* larvae under exposure to Cry3Aa toxin. The study of Cry-toxin effect on the insect metabolism is important for understanding the mechanisms of insect resistance to Cry-toxins and can contribute to the improvement of Cry-toxin preparations.

High-throughput sequencing was used to obtain EST databases from midguts of one month old *T. molitor* larvae fed either a control diet or diet containing 0.1% Cry3Aa for 24 h. Sequencing was performed by Genome Sequencer FLX System (454 Life Sciences, Roche). The reads were assembled in contigs with GS Assembler (454 Life Sciences, Roche) software.

Comparison of predicted proteomes in control and *Bt*-treated groups using BLAST2GO software was made to reveal the overall effect of the toxin on the insect. Negative effect of Cry3Aa was observed on the predicted proteins involved in catalytic activity, enzyme regulator activity, and binding molecular functions, as well as such biological processes as adhesion, localization, multi-organism and metabolic processes. Predicted proteins involved in signaling and cell components biogenesis processes were considerably increased in number.

Direct comparison of contigs expression intensity was performed based on the numbers of reads used in each contig assembly. Only those contigs that had higher than average expression in at least one EST dataset were chosen for comparison. The filtered contigs were divided in four groups: showing more than 2-fold down-regulation of expression (27 different

sequences); showing more than 2-fold up-regulation of expression (31 different sequences); presented only in control group (60 sequences); presented only in *Bt*-treated group (57 sequences). Analysis of these groups of sequences revealed:

i. Remarkable changes in digestive enzymes spectra, specifically, expression of serine peptidases mostly was down-regulated.
ii. A substantial number of ribosomal proteins was found in the contig group presented only in control database. This can testify to the decreased translation activity in toxin treated cells.
iii. Expression of hexamerin 2 beta, which is involved in oxygen transport, was decreased.
iv. Up-regulation of expression was found for apoptosis inhibitor (protein 3) and enzymes that can hydrolyze chitin.
v. Unsystematic change in expression was found for proteins involved in oxidation-reduction processes, proton transport, and oxidative stress.

The data are the first application of high-throughput sequencing to the study of *Bt* intoxication and demonstrate that Cry3Aa intoxication in *T. molitor* induces widespread changes in expression of different groups of genes.

This work was supported by the Russian Foundation for Basic Research (grants # 09-04-01449-a and 11-04-93964-SA_a). Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture.

Homologous Recombination and Horizontal Gene Transfer Play a Dominant Role in Evolution of Bacterial Genomes

Sergei Maslov

Brookhaven National Laboratory, United States, maslov@bnl.gov

I will review recent results showing that homologous recombination and horizontal gene transfer play a dominant role in evolution of bacterial genomes. For example, we recently estimated that about half of genomes of closely related E. coli strains K-12 and B underwent homologous recombination in either one strain or the other [1]. I will show how to identify and analyze such recombined regions based on specific correlation pattern of Single Nucleotide Polymorphisms in aligned genomes of bacterial strains.

I will proceed by considering another source of bacterial genome plasticity due to Horizontal Gene Transfer (HGT) of entirely new metabolic pathways from the "universal" metabolic network (metabolic core of the pan-genome of all bacterial species). Such transfers were recently proposed as an explanation of the empirical scaling law stating that in prokaryotic genomes the number of transcription factors is proportional to the square of the total number of genes. In a recent study [3] we address the question of how the topological properties of this universal network influence the power law scaling of transcriptional regulators. We also generalize the rules of our earlier model [2] to include metabolic reactions with multiple substrates and products, redundant metabolic branches and cycles, and to account for optimality of metabolic pathways. The main conclusion of our analytical and numerical modeling efforts is that the quadratic scaling holds for a broad range of universal network topologies. We also demonstrate why, in spite of the perceived "small-world" topology, real-life metabolic networks are characterized by a broad distribution of pathway lengths and regulon sizes.

[1] F W Studier, P Daegelen, R E Lenski, S. Maslov, J F Kim, "Understanding the differences between genome sequences of Escherichia coli B strains REL606 and BL21(DE3) and comparison of the E. coli B and K-12 genomes", J. Mol. Biol 394, 653-680 2009.
[2] S Maslov, S Krishna, T Y Pang, K Sneppen, "Toolbox model of evolution of prokaryotic metabolic networks and their regulation", PNAS 106, 9743-9748 2009.
[3] TY Pang, S Maslov, "Toolbox model of evolution of metabolic pathways on networks of arbitrary topology" PLoS Comp. Bio 2011 (in press)

Correlation between transcription efficiency initiation and translation efficiency for *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*

Yury MATUSHKUN, Vitaly LIKHOSHVAI, Viktor LEVITSKY
Institute of Cytology and Genetics SB RAS, Russian Federation, mat@bionet.nsc.ru

Study and optimization of genes expression efficiency are important and essential problems, as for theory so for practice. The elongation efficiency index (EEI) was developed. It has been shown [1] that there are five groups of unicellular organisms which can be separated basing on major factors affecting the genes expression efficiency at the level of translation. These factors are: codon usage bias of a gene, presence and distribution of mRNA secondary structures, "strength". Studies of the factors influence resulted in the fact that all unicellular prokaryotes and some eukaryotes can be classified in groups mentioned above. At the level of transcription, genes expression efficiency depends on, particularly, 5'-regulatory region and specifically on nucleosomes localization in it. The aim of the present study is the finding of correlation between nucleosome potential at 5'-UTR in genes of yeast species (Saccharomyces cerevisiae and Schizosaccharomyces pombe) and elongation efficiency index of the same genes. In order to characterize nucleosomes location we used the program RECON [2] calculating the nucleosome formation potential (NFP) based on dinucleotides frequencies. Verifiable hypothesis is as follows: efficient genes expression requires consistently optimized translation and transcription processes. Such a correlation was found between NFP at 5'-utr near AUG codon of yeast species (S. cerevisiae and S. pombe) genes and EEI values for the same genes.

The analysis was performed for 5649 Saccharomyces cerevisiae genes and 4546 Schizosaccharomyces pombe genes as also for 10% of genes having highest/lowest EEI values (genes were extracted from complete genomes). The sequences were extracted from gene maps of the yeasts, taken from GenBank database. EEI was calculated for all ORFs. NFP and correlation profiles constructed for regions from -600 to +600 relating to translation start.

Reliable negative correlation between NFP at 5'-UTR and EEI was found for S. cerevisiae: for all genes – in regions (-345;-230); for EEI-highest genes – in some positions at 5'-UTR (-280;-200); for EEI-lowest genes – reliable positive correlation with nucleosome potential in 5'-UTR (-250;-100) was found.

Reliable negative correlation between nucleosome potential at 5'-UTR and EEI was found for S. pombe: for all genes – (-550;-100); for EEI-highest genes – (-400;-300).

Thus selection for S. cerevisiae supported the complication of transcription initiation for low elongation rate mRNAs, by contrast, S. pombe supported transcription initiation facilitation for high elongation rate mRNAs.

It is characteristic, that correlation coefficient is positive and reliable at the region of translation start and further to 3'-end in the case of S. cerevisiae, while in the case of S. pombe it is negative and reliable at the same region. Probably, the difference is conditioned with the fact that S. cerevisiae belongs to the 1-st group of evolutionary optimization of elongation, and S. pombe belongs to the 4-th one.

The work was supported by Russian Foundation for Basic Research (No. 10-04-01310), RAS Presidium program № A.II.6, Project "Evolution of molecular-genetic systems: computer analysis and modeling" of the RAS Presidium program 25 "Biosphere origin and evolution".

1. N.V. Vladimirov, V.A. Likhoshvai, Yu.G. Matushkin (2007) Correlation of Codon Biases and Potential Secondary Structures with mRNA Translation Efficiency in Unicellular Organisms. Mol Biol (Mosk)., 41(5): 843–850.
2. Victor G. Levitsky (2004) RECON: a program for prediction of nucleosome formation potential. Nucleic Acids Res., 32: 346–349.

Splicing differences in primate brain development

Pavel Mazin[1], Mikhail Gelfand[2], Philipp Khaitovich[3]

[1] *Department of Bioengineering and Bioinformatics, Moscow State University, Vorobievy Gory 1-73, 119991, Moscow, Russia iaa.aka@gmail.com*
[2] *Institute for Information Transmission Problems RAS, Bolshoi Karetny 19, 127994, Moscow, Russia*
[3] *Partner Institute for Computational Biology CAS, 320 Yue Yang Road, 200031 Shanghai, China*

Humans differ strikingly from their close relatives, such as chimpanzees, in terms of anatomy, behavior and cognitive abilities [1]. Genetically, however, the human and the chimpanzee genomes are extremely similar. Recent studies have shown that gene expression differs substantially between humans and other primates and these differences could be linked with accelerated brain evolution on the human lineage [2]. Alternative splicing allows a single gene to produce multiple transcripts and, consequently, multiple proteins, by utilizing different splice sites in pre-mRNA. The microarrays-based analysis shown that at least 4% of expressed genes splice differently between humans and chimpanzees [3]. Here, we used next-generation sequencing technology to investigate splicing differences between humans, chimpanzees and rhesus macaques in two brain regions, prefrontal cortex and cerebellar cortex, in newborns and adults. We sequenced poly-A+ transcriptome in 30 individuals: 10 humans, 10 chimpanzees and 10 rhesus macaques, resulting in more than 15 millions reads. To reduce individual variance we pooled RNA from 5 individuals with similar age resulting in 12 samples distributed among the three species, two brain regions and two ages. We show that 2,152 genes, or 20% out of 11,008 genes expressed in brain, had significantly different splicing patterns among the three species. By contrast, only 7% and 6% of genes expressed in brain showed significant splicing differences between two brain regions and two age groups, respectively. Intriguingly, small, but substantial proportion of genes (267 genes, 2.4%) showed significant differences in age-related splicing patterns among species. These differences might contribute to developmental differences observed among these species at the level of phenotype. Our analysis of intron retention showed even more dramatic differences: 39%, 20% and 14% of genes has differences in intron retention patterns between the three species, two brain regions and two age groups, respectively. Taken together, our results indicate that splicing differences could play a profound role in the phenotypic divergence between humans and other primates.

1. G. Klein (1989) The Human Career: Human Biological and Cultural Origins, The Univ. of Chicago Press, Chicago, 1989.
2. W. Enard et al (2002) Intra- and interspecific variation in primate gene expression patterns, Science 296:340-343
3. J.A. Calarco et al (2007) Global analysis of alternative splicing differences between humans and chimpanzees, Genes Dev. 21:2963-2975

Molecular dynamics simulation of Nip7 proteins demonstrates importance of hydrophobic interaction for protein stability at high pressure

Kirill Medvedev[1], Dmitry Afonnikov[1,2]

[1]*Institute of cytology and genetics SB RAS, 10 Lavrentyeva, Novosibirsk, Russian Federation,*
[2]*Novosibirsk State University, Russian Federation*
kirill-medvedev@yandex.ru

Pressure is important environmental factor. Most of the organisms live at the atmospheric pressure ~0,1 MPa. High pressure (from 10 MPa and more) is damaging for them. However, there are piezotolerant and piezophilic organisms that live at pressures reaching several hundreds of atmospheres (tens of MPa). Analysis of proteins from these organisms will unravel mechanisms of their adaptation to high-pressure environment.

High pressure leads to denaturation of proteins as result of penetration of water molecules through pores into the hydrophobic core of the protein [1]. One of the promising approaches is molecular dynamics simulation of protein structures under increased pressure.

In this work we present results of the molecular dynamics study of high pressure influence on the structure of Nip7 protein from the hyperthermophilic Pyrococcus genus archaea of deep-sea (P.abyssi, living at 2200 m depth) and shallow-water (P. furiosus, living at 100 m depth) species. These proteins are involved in ribosomal biogenesis, participate in 27S pre-rRNA prosessing and 60S ribosomal subunit formation [2].

We investigated changes of the polypeptide chain conformation and solvent accessibility at different pressures (0.1 - 300 MPa) and temperatures (300 and 373 K). The 50 ns molecular dynamics modeling was performed using GROMACS [3].

Obtained data suggested that the RNA-binding domain of the P.abyssi Nip7 protein is more resistant to the effects of high pressure. Moreover, analysis of computer models of Nip7 proteins from P. abyssi and P. furiosus showed that the solvent-accessible surface area of proteins decreases with the pressure increasing. The area is smaller for models Nip7 P. abyssi and its relative change is less than that from P. furiosus.

In general, these data are consistent with the importance of hydrophobic interactions for the protein globule formation [4] and the presumable mechanism of destruction of protein structures under high pressure [1].

The work supported by RFBR grant 11-04-01771-a; SB RAS integration projects №109, 26, 119; REC NSU (REC-008), State contract П857 and RAS programs A.II.6 and 24.2.

1. Hummer, G., Garde, S., Garcia, A.E., et al. (1998) The pressure dependence of hydrophobic interactions is consistent with the observed pressure denaturation of proteins, Proc.Natl. Acad. Sci. USA 95:1552–1555.
2. P.P. Coltri et al. (2007) Structural insights into the interaction of the Nip7 PUA domain with polyuridine RNA, Biochemistry 46:14177-14187.
3. D.Van der Spoel et al. (2005) GROMACS: Fast, Flexible and Free. J. Comp. Chem. 26:1701-1718.
4. Kauzmann, W. (1959) Some factors in the interpretation of protein denaturation, Adv. Protein Chem. 14:1-63.

Decreased mutation rate of 5mCpG within CpG islands in the human genome

Alexander Panchin[1], Vsevolod Makeev[2], Yulia Medvedeva[2]

[1]*Department of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Russia*
[2]*Vavilov Institute of General Genetics, Russian Academy of Sciences, Russian Federation*

CpG dinucleotides are extensively underrepresented in mammalian genomes. The most common explanation for this underrepresentation is the elevated level of CpG>TpG mutations, associated with cytosine methylation in CpG dinucleotides. However, in certain genomic regions called CpG islands (CGIs), the probability of CpG>TpG mutations is much lower than in other genomic regions. This is believed to be solely due to lower level of cytosine methylation in CpG dinucleotides inside of CGIs. Using the available human embryonic stem cells methylation data, we compared 5mCpG>TpG deamination rates in human CGIs and in non-CGI genomic regions with a similar degree of methylation. To do this, we studied C/T polymorphic loci, for which the ancestral cytosine is supported by comparison with Pan troglodytes and Pongo pygmaeus. It turned out that 5mCpG>TpG transitions are less frequent within CGIs than in other genomic regions. This effect became smaller but remained significant after the local C+G and CG-content were additionally controlled for. This supports the view that the reduction of CpG>TpG transition rates in CGIs can not be solely explained by the decreased methylation rate, and that other factors are involved.

A probabilistic approach to an evolution study of sequence properties.

N. Bykova[1], R. Soldatov[2], A.A.Mironov[1,3]

[1]*Department of Bioengineering and Bioinformatics, Moscow State University, Moscow, Russia*
[2]*Department of Mechanics and mathematics, Moscow State University,Moscow, , Russia*
[3]*Institute for Information Transmission Problems (The Kharkevich Institute),Moscow, Russia.*
mironov@bioinf.fbb.msu.ru

Problem statement. Suppose we have a function of a sequence f (s): S -> R. It may be, for example, the score of the strongest TF binding site in the sequence, the energy of RNA secondary structure, the length of the tail of the animal (if we are able to calculate it from the sequence), etc. Now suppose that we have a number of related sequences and the corresponding observations of our function. The aim is to assess the probability of observation on the condition of the neutral model and thus determine the possible selection pressure on this value. If our function can be in different times under different selection pressure (e.g. under different environmental conditions) the problem is to restore the history of changes of the pressure.

To solve this problem are encouraged to use the Fokker-Planck equation. The parameters of the Fokker-Planck equation are calculated using Monte-Carlo simulations. To solve the problem of the events history reconstruction two algorithms are proposed. The algorithms are similar to the maximum probability (Viterbi) and posterior (forward-backward) decoding algorithms for hidden Markov models.

The proposed techniques were applied to the problem of evolution of signal peptides in bacterial protein families.

Acknowledgments. The research was supported by RFBR (grant #09-04-92742)

A plausible mechanism for the root apical meristem self-organization

Victoria Mironova, Ekaterina Novoselova, Nadya Omelyanchuk, Vitaly Likhoshvai
Institute of cytology and genetics SB RAS, 10 Lavrentyeva, Novosibirsk, kviki@bionet.nsc.ru

Plant architecture is formed by the activities of meristems, which comprise stem cells and their derivatives, giving rise to various cell types. The root apical meristem (RAM) is localized to the root apex. After germination RAMs of the lateral roots are formed from pericycle cells of the primary root. If the RAM was destroyed in the primary or lateral root its regeneration occurs from provascular cells. Accumulating evidence suggests that hormone auxin has the primary role in RAM patterning. The local maximum of auxin concentration forms at the precise domain in the root from the very beginning of the RAM initiation. After than it is stably maintained in development and regenerated before the RAM itself if the RAM was destroyed.

By today, the two main mechanisms of the auxin distribution formation in the root tip were proposed. The *reverse fountain* mechanism is based on a specific RAM structure in which each cell has a specified set of directions of auxin efflux [1]. The *reflected flow* mechanism is based on the auxin-dependent regulation of auxin acropetal flow: low auxin concentrations activate the transcription of *PIN1* genes, whereas the high concentrations induce degradation of PIN1 proteins [2]. We suggested that the *reverse fountain* and the *reflected flow* mechanisms are complementary in root development and their combination may provide for the RAM self-organization.

To test the hypothesis we combined both mechanisms in 2D mathematical model. This model describes (1) auxin flow from the shoot; (2) auxin synthesis that is positively regulated by auxin itself; (3) irreversible loss of auxin (degradation); (4) auxin diffusion, providing for an isotropic distribution in the root; synthesis and degradation depending on auxin concentration of (5) PIN1, (6) PIN2, (7) PIN3; (8) active auxin transport mediating by PINs proteins; (9) auxin-regulated growth and division of root cells. There are two cell types in the 2D model: central cylinder and epidermis. For the central cylinder cells the processes (1-5,7-9) are considered and described as in [2]. The processes (2-4,6-9) take place in the model epidermal cells. PIN proteins have the following locations in the cell layout: PIN1 is localized at the basal side of the central cylinder cells, PIN2 at the lateral internal and apical sides of the epidermal cells and PIN3 at all sides of potentially all cells.

For the processes (1,3-5,8-9) the parameter values were taken from [2]. Other parameters were estimated so that: (1) PIN2 is expressed predominantly in epidermal cells with low auxin level; (2) PIN3 expression domain is localized in the zone of high auxin level; (3) auxin synthesis rates are high in the cells with high auxin level. With this set of parameters and initial uniform auxin distribution, the model provides steady-state auxin distribution pattern that agree well with the experimental data. The auxin pattern established in the model solutions can be interpreted in terms of the theory of positional information specifying main cell types in the root tip. The local auxin concentrations generated *in silico* make it possible to calculate the rates of root cell division and compare the resulting dynamic characteristics (cell position, auxin concentration, and division rates) with the characteristics of cell types located *in vivo* in the root tip. Thus, the numerical simulations showed that the RAM self-organization can be predetermined by the auxin distribution.

We showed *in silico* that the 2D model reveals both the robustness to the developmental processes from the *reverse fountain* mechanism [1] and the plasticity to the environmental changes from the *reflected flow* mechanism [2]. Thus, the 2D model of auxin distribution in root provides enough wide range of possibilities to be a powerful tool for investigation of root development *in silico*.

The work is partially supported by the RAS programs № A.II.5.26, A.II.6.8, B.27.29, SB RAS 107, 119, and RFBR 10-01-00717-a,11-04-01254-a.

1. VA Grieneisen et al. (2007) Auxin transport is sufficient to generate a maximum and gradient guiding root growth, Nature, **449**:1008-1013.
2. VV Mironova et al., (2010) A plausible mechanism for auxin patterning along the developing root, *BMC Systems Biology*, **4**:98.

HETEROGENEITY OF INTERNAL TRANSCRIBED SPACERS (ITS) OF RIBOSOMAL RNA OPERON IN THE GENOMES OF TERMITES FROM CENTRAL ASIA

Gulnara S. Mirzaeva[1], Rustamjon Kh. Allaberdiev[2], Aloviddin Sh. Khamraev[1]
Kirill V. Mikhailov[3], Vladimir V. Aleoshin[3]

[1]*Uzbek Academy of Sciences, Institute of Zoology, 1, A.Niyazov, Tashkent 100095, Republic of Uzbekistan, m_Gulnora@rambler.ru*
[2] *Uzbek Academy of Sciences, Scientific Centre of Plant Production "Botanika", 32 F.Khodjaev, Tashkent 100125, Republic of Uzbekistan,*
[3]*Belozersky Institute for Physicochemical Biology, Lomonosov Moscow State University, Moscow 119991, Russian Federation, Aleshin@genebee.msu.su*

Anacanthotermes turkestanicus and *Anacanthotermes ahngerianus* are two morphologically highly similar species of termites of genus *Anacanthotermes* that are widespread in Central Asia. These two species of termites leave significant impact on the industry and infrastructure in the region by infesting and destroying wooden constructs. Effective application of pest control methods depends on successful species identification but is complicated by their high degree of similarity. Here we have attempted to use the sequences of internal transcribed spacers of rRNA operon to reliably differentiate between the two species. The ITS sequences of rRNA operon, ITS1 and ITS2, are frequently used in phylogenetic analysis at low taxonomic level due to their higher rate of variability and ease of obtainment: the two spacers are sandwiched between the conserved genes for structural rRNAs and are usually present in multiple copies in the genome. It is also common to observe little to no variation between their copies within the genome, which is a result of concerted evolution of rRNA operon [1, 2]. For this work we obtained gene fragments containing both ITS sequences from individual termites collected from three regions in Uzbekistan: Samarkand, Bukhara and Khorezm. We found that in each population individual termites contain multiple clearly distinct variants of ITS sequences, thus constituting a relatively rare exception to the homogenizing effect of concerted evolution [3-5]. Some of the ITS variants differ considerably, containing multiple substitutions, large indels, and repeat expansions. Further, the difference between ITS variants in an individual often exceeds that of similar ITS variants from different populations. This result indicates that the ITS sequences of *Anacanthotermes* diverged within their genomes prior to the settlement of these termites on the territory of Uzbekistan.

This work was supported by the Russian Foundation for Basic Research.

1. G.Dover (1982) Molecular drive: a cohesive mode of species evolution, *Nature,* **299:**111–117.
2. C.Polanco et al. (1998) Multigene family of ribosomal DNA in *Drosophila melanogaster* reveals contrasting patterns of homogenization for IGS and ITS spacer regions. A possible mechanism to resolve this paradox, *Genetics,* **149:**243–256.
3. J.H.Gunderson et al. (1987) Structurally distinct, stage-specific ribosomes occur in *Plasmodium, Science,* **238:**933–937.
4. S.Carranza (1996) Evidence that two types of 18S rDNA coexist in the genome of *Dugesia* (Schmidtea) mediterranea (Platyhelminthes, Turbellaria, Tricladida), *Journal of Molecular Evolution,* **13:**824–832.
5. M.J.Telford, P.W.H.Holland (1997) Evolution of 28S ribosomal DNA in chaetognaths, duplicate genes and molecular phylogeny, *Journal of Molecular Evolution,* **44:**135–144.

Computational studies for understanding mechanistic details of newly discovered Post-Translational Modifications

Shradha Khatar and Debasisa Mohanty

Bioinformatics Center, National Institute of Immunology, New Delhi, India
deb@nii.res.in

Post Translational Modification (PTM) typically involve covalent attachment of chemical groups to amino acids of a protein and they play a central role in intracellular signaling in both eukaryotic and prokaryotic cells. Because of their role in central signaling pathways they are either stimulated or targeted by pathogens for promoting their survival in the host. Therefore for identification of enzymes involved in novel PTMs and their target proteins/pathways has been an active area of research. In this work, we have carried out computational studies to analyze various enzymes and target proteins involved in two of the newly discovered post translational modifications, namely AMPylation and Eliminylation.

Ampylation is the enzymatic transfer of AMP moiety from ATP, to tyrosine/threonine/serine residues of eukaryotic substrate proteins (1). AMPylation is catalyzed by Fic (Filamentation induced by cAMP) domains which are present in many bacterial pathogens as well as higher eukaryotes including humans. Recent experimental studies indicate that AMPylation could also be catalyzed by GS-ATase (glutamine synthetase adenylyl transferase) domains present in effector protein DrrA from human pathogen Legionella pneumophila (3). Even though known substrates of AMPylation domains are RhoGTPases, based on indirect evidence it has been suggested that proteins other than RhoGTPases could also be AMPylated (1). This suggests that AMPylation might play a major role in variety of cellular processes involving different families of AMPylating domains and their substrate proteins. However, deampylators that would reverse the effect of AMPylation bought about by Fic domains have not been characterized yet. We have carried out comprehensive computational analysis involving

profile HMM and SVM approaches for identifying new AMPylation domains from among the unannotated proteins in genomes of various organisms. We have carried out molecular dynamics (MD) simulations on crystal structure of Fic-cdc42 complex to understand the structural basis of substrate selection by AMPylation domains. We have also analyzed the known biochemical reactions catalyzed by various folds to identify other catalytic domains which might potentially carry out AMPylation and deampylation reactions.

Eliminylation is another novel PTM which is mediated by newly discovered phosphothreonine lyase enzymatic domains in different bacterial effector families (3). Phosphothreonine lyase irreversibly converts phosphothreonine into dehydrobutyrine, which cannot be phosphorylated again. Such PTMs involving phosphothreonine lyase activity have also been attributed biosynthesis of lantibiotics (4). We have carried out structure as well as profile based search for identifying phosphothreonine lyase catalytic domains in various organisms and analyzed their possible biological functions.

1. Yarbrough, M. L., Li, Y., Kinch, L. N., Grishin, N. V., Ball, H. L. and Orth, K. (2009) AMPylation of Rho GTPases by Vibrio VopS disrupts effector binding and downstream signaling. Science, 323, 269-272.
2. Muller, M.P., Peters, H., Blumer, J., Blankenfeldt, W., Goody, R.S. and Itzen, A. (2010) The Legionella effector protein DrrA AMPylates the membrane traffic regulator Rab1b. Science, 329, 946-949.
3. Li, H., Xu, H., Zhou, Y., Zhang, J., Long, C., Li, S., Chen, S., Zhou, J.M, and Shou, F. (2007). The phosphothreonine lyase activity of a bacterial type III effector family. Science 315, 1000-1003.
4. Goto Y, Li B, Claesen J, Shi Y, Bibb MJ, van der Donk WA. Discovery of unique lanthionine synthetases reveals new mechanistic and evolutionary insights (2010) PLoS Biol. 8:e1000339.

Homology-based modeling of 3D Structure of Human Alpha-fetoprotein in Complex with Estrogens

Alexander Terentiev[1], Nurbubu Moldogazieva[1], Olga Levtsova[2], Denis Borozdenko[1], Dmitry Maximenko[1], K. Shaitan[2]

[1]*Russian State Medical University, Ostrovityanova street, 1, Moscow, Russia,* *aaterent@inbox.ru*, *nmoldogazieva@mail.ru*
[2]*Moscow State University, Russia, Vorobyovy Gory, 1,* *shaitan@moldyn.org*

Introduction. Alpha-fetoprotein (AFP) is a major mammalian oncofetal protein, which is also present in small quantities (about 10ng/ml) in adults. AFP is a glycoprotein with M.W. of 68–73kDa and carbohydrate content of 3–5%. Primary structure of human AFP is represented by 609 aa residues [1–2]. AFP belongs to the family of proteins – products of albuminoid genes, which are located in tandem arrangement in chromosome 4 (region 4q11–q13). This family includes, except AFP itself, also albumin, alpha-albumin (afamin) and vitamin-D-binding protein (VTDB). Proteins of this family demonstrate considerable similarity of their primary and secondary structures. Pairwise alignment of amino acid sequences shows that identity between human AFP and albumin is 40%, whereas identity between human AFP and VDTB is 16% (this is because of absence of C-terminal fragment in VTDB).

Ability to bind free estrogens has been demonstrated for rodent AFPs, but not for human AFP. Using affinity chromatography it was shown that human AFP is able to bind immobilized diethylstilbestrol (DES) and, in less extent, estradiol-17beta [3–5].

Difficulties in crystallization of AFP makes it impossible to study its 3D structure by experimental methods such as X-ray crystallography or/and nuclear magnetic resonance (NMR). Also, molecular mechanisms of AFP interaction with estrogens remain undiscovered.

Aims. The goal of this study was to model 3D structure of human AFP on the basis of homology with serum albumin and VTDB and to investigate molecular mechanisms underlying its interaction with estrogens.

Methods. Crystal structures of albumin and VTDB are obtained experimentally and are given in the Protein Data Bank (PDB). Multiple alignment of amino acid sequences of AFP, albumin and VTDB was perform using ClustalW program. 3D model of human AFP was made using MODELLER program. Docking of DES and estradiol-17beta to AFP molecule was performed using the program Autodock. To perform relaxation and

optimization of obtained complexes molecular dynamic simulation by GROMACS software package was used.

Results. In PDB we found 68 3D structures of human serum albumin and six 3D structures of VTDB. Structures with no ligand - two for albumin (1AO6 and 1E7A) and one for VTDB were selected as templates for modeling. We obtained 3D structure of AFP in which N-terminus (aa 1–28) was unstructured, because our templates do not demonstrate homology in this region. RMSD for the whole AFP molecule and its estrogen-binding site were calculated to estimate the stability of the structure.

It was revealed that estrogen-binding interface is located in a cavity formed in human AFP molecule. The alpha-helix, responsible for hormone binding, is on the bottom of the cavity. 5ns MD simulation of the complex showed that estrogens interact with binding site through formation of hydrogen bonds and hydrophobic interactions. Interaction between the hormones and the binding site is provided by 1) hydrogen bonding with participation of OH-group of hormones and polar atoms of aa side chains (R452, E551 and S445 in case of DES) and 2) hydrophobic interactions with involvement of aromatic ring of hormones and hydrophobic aa side chains (between L138, I450 and M448 in DES).

Conclusions. Model of 3D structure of AFP was constructed on the basis of homology and estrogen-binding site on the bottom of a cavity was revealed. This model provides rationale for understanding molecular mechanisms of binding of estrogens to human AFP.

References

1. A.J.Luft, F.L. Loscheider (1983) Structural analysis of human and bovine α-fetoprotein by electron microscopy, image processing and circular dichroism. *Biochemistry*, **22:**5971–5978.
2. A.A. Terentiev, N.T. Moldogazieva (2006) Structural and functional mapping of human alpha-fetoprotein. *Biochem. (Mosc.)*, **71:**120–132.
3. G.J. Mizejewski (2004) Biological roles of alpha-fetoprotein during pregnancy and prenatal development. *Exp. Biol. Med.*, **229:**439–463.
4. J. Uriel, B. de Nechaut, M. Dupiers (1972) Estrogen-binding properties of rat, mouse and man fetospecific serum protein. Demonstration by immuno-autoradiographic methods. *Biochem. Biophys. Res. Commun.*, **46:**1175–1180.
5. Yu. S. Tatarinov, A.A. Terentiev, N.T. Moldogazieva, A.K. Tagirova (1991) Human alpha-fetoprotein and its purification by chromatography on immobilized estrogens. *Tumor Biol.*, **12:**125–130.

Thiosemicarbazone Derivatives as Potent RNR Inhibitors: *In Silico* Based Pharmacophore, Binding Mode and Toxicity Analysis

N.S. Hari Narayana Moorthy, Nuno Cerqueira, Maria Ramos and Pedro Fernandes;

REQUIMTE, Departamento de Química, Faculdade de Ciências, Universidade do Porto, 687, Rua do Campo Alegre, 4169-007, Porto, Portugal, hari.moorthy@fc.up.pt

In silico based pharmacophore distance study, binding mode analysis and toxicity study were performed on a series of α-N-heterocyclic carboxaldehyde thiosemicarbazones derivatives exhibit anticancer activity by inhibiting ribonucleotide reductase (RNR) enzyme. The flexialigned structure of the active compounds in the series was used as a pharmacophore query structure (template) (Fig 1). The active conformers obtained from the pharmacophore based conformational search method shows that Aro/Hyd, Acc and Don properties (pharmacophore contour) of the compounds are important for the interaction and RNR inhibitory activity. It also revealed that the distance between the Aro/Hyd contours on both side of the Acc/Don (centered pharmacophore contour) (bridged groups) are important for the inhibitory activity and most of the conformers selected from the active compounds in the series have the distance (Aro/Hyd) (F3 & F4) is 8.13-8.50 Å. The pharmacophore distance between the acceptor site (F5) and the Aro/Hyd contour (F3) is 3.72 Å for active compounds and the less active/inactive compounds have the distance of 4.18, 3.63, 3.70 and 4.28 Å (Fig. 2). The physicochemical descriptors calculated for the active conformers obtained from the pharmacophore analysis were used for the structural feature analysis for ribonucleotide reductase inhibitory activity, hERG blocking activity and toxicity of the compounds. These results show that the flexibility of the acceptor/donor groups in the bridge and increased VDW surface area with hydrophobic properties of those compounds is important to improve enzyme ligand interaction. Some hydrophobic properties on the vdW surface of the molecules are favorable for the hERG blocking and toxicity of the compounds, and are unfavorable for the RNR inhibitory activity. This study will helpful for further free energy analysis and inhibitor design for ribonuceotide reductase enzyme.

One of the Authors (N.S.H.N. Moorthy) grateful by acknowledges the Foundation of Science and Technology (FCT), Portugal for Postdoctoral Grant (SFRH/BPD/44469/2008).

1. N.M. Cerqueira, et al. (2005) Overview of ribonucleotide reductase inhibitors: An appealing target in anti-tumor therapy. *Cur. Med. Chem.*, 12:1283-1294.
2. N.S.H.N. Moorthy, et al. (2011) QSAR analysis of 2-benzoxazolyl hydrazone derivatives for anticancer activity

and its possible target prediction. *Med. Chem. Res.*, DOI: 10.1007/s00044-010-9510-3.
3. N.M. Cerqueira, et al. (2007) Ribonucleotide reductase: A critical enzyme for cancer chemotherapy and antiviral agents. *Recent Pat. Anticancer Drug Dis.*, **2:**11-29.

Figure 1: *Pharmacophore query structure (template) flexialigned compounds*

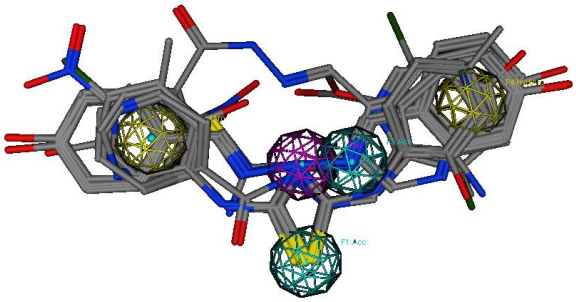
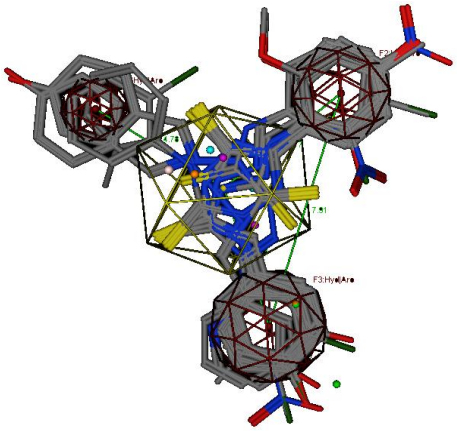


Figure 2: *Flexialigned structure of the conformers selected by pharmacophore based conformation search approach*



Binding Feature Analysis of hERG Blockers: A Computational Study

N.S. Hari Narayana Moorthy, Nuno S Cerqueira
Dept. of Chemistry and Biochemistry, University of Porto, Porto, Portugal, hari.moorthy@fc.up.pt

The human ether-a-go-go-related gene (hERG) encodes the major protein underlying the rapid delayed rectifier K+ current in the heart and it causes long QT syndrome when blocking. Many marketed drugs such as sertindole, grepafloxacin, terfenadine, etc were withdrawn due to this effect. Hence, we are concentrating to develop novel bioactive moiety without this unwanted side effect and with desired ADME properties. In the present investigation, we have performed QSAR and pharmacophore analysis on some structurally different hERG blockers to determine the kind of molecular feature responsible for the hERG blocking action. The QSAR analysis results show that the van der Waals volume and surface properties play important role for the blocking action. The vsurf_HB descriptors reveals that the hydrogen bonding capacity of the molecule should be significant and the hydrophilic-lipophilic (vsurf_HL) properties should be balanced for significant action. The vsurf_CW, a capacity factor reveals that the hydrophilicity of the molecules on unit surface area should be low and there should be a balance between hydrophilic and hydrophobic properties. To support this statement, the other properties such as the integy moments (vsurf_IW and vsurf_ID) of the molecules suggest that there should be a clear separation between the hydrophilic and hydrophobic region and the hydrophobic area should be concentrated in particular region. The pharmacophore analysis on the compounds also explains that the polar properties are distributed in the molecular surface and the aromatic/hydrophobic property of the molecules should be far away for the polar site. The results reveal that the active site of the protein should have some polar environment for the hydrogen bonding interaction and should be away from the hydrophobic interaction sites. The results derived from this study along with the ongoing work in our laboratory will be helpful for designing novel moieties with less/free of hERG blockade.

Mathematical model of the inhibiting part in TCA at Citric Acid synthesis by superproducers cross-mutants of Yarrowia lipolytica from glucose

Yulia Lunina[1], Andrew Rudenko[2], Igor Morgunov[1]

[1]*G. K. Skryabin Institute of Biochemistry and Physiology of Microorganisms Russian Academy of Sciences prospect Nauki 5, Pushchino Moscow Region 142290, Russian Federation, morgunovs@rambler.ru*
[2]*Faculty of Physics, Lomonosov Moscow State University, 1, building 2, GSP-1, Leninskiye Gory, Moscow, 119991 Russian Federation*

Abstract. As described in (1) with processing of natural yeast *Yarrowia lipolytica* 704 by UV irradiation and mutagen N-methyl-N'-nitro-N-nitrosoguanidin (NG), 1500 variants were obtained, and three of these mutant strains were found to be an excellent citric acid (CA) superproducers on glucose. Acid-formatting activity of one of those mutants (N 15) essentially (on 50 %) exceeded similar parameter of the natural strain. Furthermore, the mutant mass yield was 43.0 % from consumed glucose, and considerably exceeded that of natural strain (19%). Natural strain and mutant N 15, cultivated in fermenters, both accumulated CA at significant levels, about 70 g/l. But the mutant strain produced this amount for 3 day of cultivation, and natural yeast accumulated the same amount only on 5 day of cultivation. Optimization of nutrient media and cultivation parameters has not helped essentially to overcome the barrier of 70 g/l, neither at the natural strain nor at the mutant. Apparently, it is connected with the inhibition of some metabolic reactions, involved in CA oversynthesis. To understand, at what synthesis stage the inhibition occurs, it was necessary to construct a mathematical model.

For a basis the Wayman and Tseng equation with Andrews type inhibition was taken (2). The Andrews function was combined with a linearly decreased activity function to form a five-parameter discontinuous model.

Model fitting. The parameter values obtained correspond with general concepts of CA oversynthesis by yeast and the literature data (3, 4, 5).

Results

1. CA yield levels of about 70 g/l were reached stably were slightly overcame;

2. The exceeding of the levels were statistically significant (σ^2 and Least Square Method) but small in size (some percents);

3. The system's behavior is non-conservative, no stability on small parameter was found, the inhibiting starts abruptly and stochastically.

As the future prospect: to model the whole TCA cycle not only from point of view on searching the "bottlenecks", but to create the full mathematic model of all chemical changes in TCA - on analogy with well investigated and modeled thermodynamic Carno cycle.

Acknowledgements. The authors are very grateful to doctor Svetlana V. Kamzolova (Institute of Biochemistry and Physiology of Microorganisms Russian Academy of Sciences, Pushchino) for the numerous valuable remarks made during discussion of given clause.

We are searching for the young scientists who are not afraid of terms entropy, enthalpy etc.

1. T.V.Finogenova et al. (2008) Obtaining of the mutant Yarrowia lipolytica strains producing citric acid from glucose, *Prikl. Biokhimiya i Microbiologiya* 44, **2:** 219-224. (rus)
2. G.Alagappan, R.M.Cowan (2001) Biokinetic models for representing the complete inhibition of microbial activity, Biotechnology and bioengineering 75, 4: 393-405.
3. L.M.Glazunova, T.V.Finogenova (1976) Enzyme activity of citrate, glyoxylate and pentose phosphate cycles during synthesis of citric acids by Candida lipolytica, Microbiologiya XLV, vol. 3: 444-449. (rus)
4. I.T.Ermakova, T.V.Finogenova (1971) Participation of glyoxylate cycle in metabolism of alkane-oxidizing yeast Candida lipolytica during biosynthesis of α-keto-glutaric acid, Microbiologiya XL, vol. 2: 223-226. (rus)
5. I.G.Morgunov et al. (2004) Regulation of NAD+-dependent isocitrate dehydrogenase in the citrate producing yeast Yarrowia lipolytica, Biochemistry (Moscow) 69, 12: 1391-1398.

Evolution of membrane bioenergetics

Daria V. Dibrova[1,2], Michael Y. Galperin[3], Armen Y. Mulkidjanian[1,4]

[1]*School of Physics, University of Osnabrueck, D-49069 Osnabrueck, Germany;*
[2]*School of Bioengineering and Bioinformatics, Moscow State University, Moscow 119992, Russia;*
[3]*National Center for Biotechnology Information, NLM, NIH, Bethesda, MD 20894, USA*
[4]*A.N.Belozersky Institute of Physico-Chemical Biology, Moscow State University, Moscow, 119991, Russia*.

By combining structural and phylogenetic analyses, we have earlier clarified the evolutionary relationships among membrane enzymes that couple the transmembrane transfer of protons or sodium ions with the synthesis/hydrolysis of ATP. A comparison of the structures of the sodium-dependent bacterial and archaeal ATPases revealed nearly identical sets of amino acids involved in sodium binding. Phylogenetic analysis showed that the sodium-dependent ATPases are scattered among proton-dependent ATPases in both the archaeal and bacterial branches of the phylogenetic tree [1, 2]. Barring convergent emergence of the same set of amino acid ligands in several lineages, these findings indicate that the use of sodium gradient for ATP synthesis is the ancestral modality of membrane bioenergetics and that the Last Universal Common Ancestor (LUCA) was unlikely to have proton-dependent energetics [3].

A focused search for primitive sodium-translocating ATPases/ATP synthases among microbial genomes identified an atypical form of the F_1F_O-type ATPase that is encoded in a number of phylogenetically diverse marine, halotolerant and pathogenic bacteria and in the archaea *Methanosarcina barkeri* and *M. acetivo*rans [4]. In complete genomes, representatives of this form (referred to here as an N-ATPase) are always present as second copies, in addition to the typical proton-translocating ATP synthases. The N-ATPase is encoded by a highly conserved operon and its subunits cluster separately from the equivalent subunits of the typical F-type

ATPases. Other distinctive properties of the N-ATPase operons include the absence of the *delta* subunit from its cytoplasmic sector and the presence of two additional membrane subunits, AtpQ (formerly gene 1) and AtpR (formerly gene X). The N-ATPases carry a full set of sodium-binding residues, indicating that most of these enzymes are sodium-translocating ATPases that likely confer on their hosts the ability to extrude sodium ions.

The evolutionary primacy of the sodium-dependent bioenergetics contradicts the common belief that the LUCA possessed several proton pumps, such as cytochrome oxidase and quinol:cytochrome c oxidoreductase, which are widespread among bacteria and archaea. To address this conundrum, we analysed the phylogeny of the quinol:cytochrome c oxidoreductases and showed that the phylogenetic tree of quinol:cytochrome c oxidoreductases did not follow the 16S RNA tree. We hypothesize that the common ancestor of the quinol:cytochrome c oxidoreductases evolved within bacteria, perhaps in response to the emergence of chlorophyll-based photosynthesis. Different archaeal phyla seem to have acquired different types of quinol:cytochrome c oxidoreductases from bacteria by lateral gene transfer on several independent occasions. A similar scenario has been recently proposed for the evolution of the cytochrome oxidases [5].

Generally, the evolution of membrane bioenergetics seems to have followed the overall trend of progressive sequestration of protocells from the environment [6]. The first ATP synthases likely evolved from protein translocases [7] within primitive, sodium-impermeable but proton-permeable cell membranes that harboured a set of sodium-transporting enzymes. The more structurally demanding proton-tight membranes, which could accommodate proton-specific pumps, appear to emerge later, independently in bacteria and archaea.

Acknowledgements: The authors appreciate the support from the *Deutscher Akademischer Austausch Dienst*, *Deutsche Forschungsgemeinschaft*, the Russian Foundation for Basic Research (10-04-91331) and the Russian Government (02.740.11.5228).

References

1. A.Y. Mulkidjanian, P. Dibrov, M.Y. Galperin (2008) The past and present of sodium energetics: may the sodium-motive force be with you. *Biochim Biophys Acta* 1777:985-992.
2. A.Y. Mulkidjanian, M.Y. Galperin, K.S. Makarova, Y.I. Wolf, E.V. Koonin (2008) Evolutionary primacy of sodium bioenergetics. *Biol Direct* 2008, 3:13.
3. A.Y. Mulkidjanian, M.Y. Galperin, E.V. Koonin (2009) Co-evolution of primordial membranes and membrane proteins. *Trends Biochem Sci*, 34:206-215.
4. D.V. Dibrova DV, M.Y. Galperin, A.Y. Mulkidjanian (2010) Characterization of the N-ATPase, a distinct, laterally transferred Na^+-translocating form of the bacterial F-type membrane ATPase. *Bioinformatics* 26:1473-1476.
5. J. Hemp, R.B. Gennis (2008) Diversity of the heme-copper superfamily in archaea: insights from genomics and structural modeling. *Results Probl Cell Differ*, 45:1-31.

The fitness conferred by recently replaced amino acids rapidly declines with time

Sergey Naumenko[1,2], Georgii Bazykin[1,2], Alexey Kondrashov[1,3]

[1]*M. V. Lomonosov Moscow State University;Russian Federation*
[2]*Institute for Information Transmission Problems, Russian Academy of Sciences, Russian Federation*
[3]*University of Michigan, USA*
sergey.naumenko@yahoo.com

Fitness landscape, the map that relates fitnesses to genotypes, can change in the course of evolution. One can expect the relative fitness of the allele that is currently fixed at a locus to increase, because adaptive allele replacements at other loci always increase its fitness, but not necessarily fitnesses of other, currently absent, alleles. In contrast, little is known about evolution of fitnesses of currently absent alleles. Here, we show that fitness of a recently replaced allele declines disproportionally rapidly. At a protein site, soon after an A > B amino acid replacement, B > A reversals occur 3.7 times more often than B > C replacements, where C is any amino acid, different from both the ancestral A and the derived B. However, after a time interval sufficient for accumulation of ~0.6 synonymous substitution per site, B > A reversals occur only 1.8 times faster than B > C replacements. This pattern can be explained only by a rapid, specific decline of the fitness conferred by A during a relatively short period of time after its replacement. This effect demonstrates that many amino acid replacements which are not completely selectively neutral, nevertheless do not increase fitness and, instead, only chip away unused parts of the fitness landscape, due to pervasive epistasis.

Interactions of antimicrobial peptide buforin 2 with nucleic acids: how P11A-substitution modulates structure and functioning

Tatsina Naumenkova

Lomonosov Moscow State University, Faculty of Biology, Russian Federation, tnaumenkova@gmail.com

Buforin 2 is one of the most effective amongst known antimicrobial peptides. It belongs to abundant group of membrane-active peptides. However, it is hypothesized to kill microorganisms by entering cells and binding nucleic acids. Ability to disrupt membrane structures is supposed to be closely interconnected with alpha-helical structure of antimicrobial peptides. In buforin 2 alpha-helix is distorted by P11 residue in the middle of the molecule.

Earlier we showed that P11A-substituted analogue of buforin 2 bound model membrane of prokaryotic cells. Group of four analogue molecules disintegrated membrane structure. Toroidal pore assembled of four analogue molecules was more stable than that consisted of four molecules of native peptide.

In this study we assessed the ability of buforin 2 and its analogue to bind nucleic acids.

Simulations were run in OPLS-AA force field using Gromacs 3.3.2 package. Initial structures of both peptides were created in HyperChem 7.5. They had alpha-helical secondary structure; N- and C-termini were ionized. We used TIP4P water model and added Na+ and Cl- ions to neutralize net charge of the systems. Before all simulations, systems energy was minimized over 10000 steps using steepest descent algorithm. MD simulations used stochastic dynamics. The cutoff radii for electrostatic and Van der Waals interactions were 2 nm; temperature was maintained at 300 K. All systems were first relaxed for 100 ps.

We observed both peptides bind DNA segment. Before binding, the peptides formed curvature in their structures. The curvature increased amphipathic properties of the peptides. We analysed secondary structure of the peptides using DSSP standard procedure. Secondary structure of the peptides bound to DNA was different from that observed for peptides bound to membrane surface. Total energy of the systems, electrostatic and Van der Waals terms were assessed. The amount of hydrogen bonds of different types in the systems was also considered.

Our results suggest that P11A-substituted analogue of buforin 2 is not only selective membrane-active peptide but also possesses initial antimicrobial properties of native peptide.

The author is grateful to professor K.V. Shaitan for strict and intensive supervision and useful discussions.

Hierarchical classification of glycoside hydrolases

Daniil Naumoff

S. N. Winogradsky Institute of Microbiology, Russian Academy of Sciences, Prospekt 60-letiya Oktyabrya 7/2, Moscow 117312, Russia; State Institute for Genetics and Selection of Industrial Microorganisms, I-Dorozhny proezd, 1, Moscow 117545, Russia; daniil_naumoff@yahoo.com

Glycoside hydrolases or glycosidases (EC 3.2.1) are a widespread group of enzymes hydrolyzing various carbohydrates and glycoconjugates. They are represented in almost all living organisms. Their catalytic domains are grouped into 120 sequence-based families in the CAZy database (http://www.cazy.org/): GH1–GH125, except GH21, GH40, GH41, GH60, and GH69. 51 of these families compose 14 clans (GH-A–GH-N) at a higher hierarchical level. Enzymes of the same clan have common evolutionary origin of their genes and share the most important functional characteristics: composition of the active center, anomeric configuration of the hydrolyzed glycosidic bond, and molecular mechanism of the catalyzed reaction (either inverting, or retaining). The subfamily level of the classification exists only for GH13 and GH30 families in the CAZy database. However, extensive data on relationship between glycosidase families belonging to different clans and/or non-included into any clans are available in the literature, as well as information on phylogenetic protein relationship within particular families. Based on this information and on our complementary data we propose a multilevel hierarchical classification of glycosidases and their homologues.

According to the SCOP database (http://scop.mrc-lmb.cam.ac.uk/scop/), the catalytic domains of almost all families of glycoside hydrolases have one of six basic folds: β/α-barrel, β-propeller, β-jelly roll, α/α-barrel, β-solenoid, and lysozyme-type. Extensive comparative analysis of the primary and tertiary protein structures suggests the common evolutionary origin of essentially all glycosidase domains having the same type of the three-dimensional structure. This fact allows us to classify glycoside hydrolases into six main primary groups based on homology of their catalytic domains. Grouping of glycoside hydrolase families into clans at the Pfam database (http://pfam.sanger.ac.uk/) suggests the subsequent level in the hierarchical classification of glycosidases. Clans from the CAZy database represent a lower level, combining together the closest families. Useful information about homology-based grouping of glycosidases can be obtained from many other on-line protein classifications, including COG, KOG, PANTHER, and Génolevures. Intrafamily phylogenetic analysis of proteins suggests the lowest classification level – the subfamily level. Iterative screening of the protein database by PSI-BLAST program results in finding of interfamily relationships and, in some cases, allows to

distinguish additional intermediate level(s) in the classification. Evolutionary connections of glycosidase families with other families, including families of functionally uncharacterized domains, also can be traced by the iterative screening of the database. The obtained results allow to extend the hierarchical classification of glycoside hydrolases on their homologues.

The TIM-barrel or β/α-barrel is the most common protein fold among enzymes. About a half of glycoside hydrolase families has catalytic domain with this type of the three-dimensional structure. The majority of them have the classical $(\beta/\alpha)_8$-barrel fold and belong to "(Trans)glycosidases superfamily" in SCOP and to "TIM barrel glycosyl hydrolase superfamily" (or clan CL0058) in Pfam databases. According to the CAZy database, some of the corresponding families form four clans: GH-A, GH-D, GH-H, and GH-K. We distinguish two additional hierarchical levels – "type" and "superfamily" – between Pfam- and CAZy-clans, which we name "fold" and "clan" levels respectively. Currently, 46 CAZy-families (GH1, GH2, GH3, GH5, GH10, GH13, GH14, GH17, GH18, GH20, GH25, GH26, GH27, GH29, GH30, GH31, GH35, GH36, GH39, GH42, GH44, GH50, GH51, GH53, GH56, GH59, GH66, GH67, GH70, GH71, GH72, GH77, GH79, GH84, GH85, GH86, GH89, GH97, GH98, GH99, GH101, GH107, GH112, GH113, GH114, and GH123), as well as, 55 families of functionally uncharacterized proteins (GHL1–GHL50, COG1306, COG1649, COG2342, PF11308, and PF11790) are combined into the classical $(\beta/\alpha)_8$-barrel fold. Four additional families – GH6, GH38, GH57, and GH119 – belong to the unusual $(\beta/\alpha)_7$-barrel fold. These two folds are most probably evolutionary related and compose one of the main primary groups in our classification – the β/α-barrel group of folds. Families GH13 (clan GH-H) and GH36 (GH-D) are heterogeneous groups of proteins and they can be divided into respectively three and eleven smaller families with evidently monophyletic status.

COG2342 is a family of hypothetical glycoside hydrolases

Daniil Naumoff[1,2], Olga Stepuschenko[3]

[1]*S. N. Winogradsky Institute of Microbiology, Russian Academy of Sciences, Moscow, Russia*
[2] *State Institute for Genetics and Selection of Industrial Microorganisms, Moscow Russia*
daniil_naumoff@yahoo.com
[3]*Kazan Federal University, Kazan, Russia*

COG2342 is a family of prokaryotic protein domains. As it was shown for *Thermotoga maritima* protein (PDB, 2AAM), domains of COG2342 family have the TIM-barrel type of the three-dimensional structure. It is known for many years [1] that this family is closely related to GH114 family of glycoside hydrolases (or COG3868). GH114 is a small family of poorly characterized proteins: the endo-alpha-1,4-polygalactosaminidase activity (EC 3.2.1.109) has been shown only for two representatives of the family. Iterative screening of the protein database, using GH13 and GH31 domains as a query, allowed us to reveal their evolutionary connections with COG2342 [2] and COG3868 [3] families, respectively.

We have revealed 234 non-identical protein sequences of COG2342 domains using the blast algorithm. They include representatives of Archaea (Crenarchaeota, Euryarchaeota, and Korarchaeota) and Bacteria (Actinobacteria, Aquificae, Bacteroidetes, Deinococcus, Dictyoglomi, Firmicutes, Fusobacteria, Lentisphaerae, Nitrospirae, Planctomycetes, Proteobacteria, Spirochaetes, and Thermotogae). Two protein fragments (GenPept, EEE74975.1 and EEE71697.1) deposited in the database were obtained in the frame of black cottonwood *Populus trichocarpa* genome project. However, we assume that they are originated from a bacterial contamination of the plant tissue. Two highly conserved residues (Asp and Glu) are most probably the nucleophile and proton donor in the active center of COG2342 hypothetical glycoside hydrolases, respectively. We used members of GH114 family as outgroup for the phylogenetic analysis. Topology of the maximum parsimony and neighbor-joining trees did not allow us to distinguish clear subfamilies in COG2342 family. The important role of horizontal transfer in the evolution of COG2342 and COG3868 proteins can be suggested.

Iterative screening of the protein database by PSI-BLAST allowed us to reveal relationship of COG2342 with GH5, GH13, GH18, GH20, GH27, GH29, GH31, GH35, GH36A, GH36B, GH36F, GH36G, GH36H, GH36J, GH36K, GH42, GH66, GH72, GH97, GH101, GH114, COG1306, COG1649, GHL3, GHL4, GHL5, GHL7, GHL9, GHL10, GHL11, GHL13, GHL14, GHL30, GHL39, GHL48, GHL49, and CE9. These data support the common evolution origin of all TIM-barrel type glycoside hydrolase catalytic domains, as we suggested earlier [4, 5].

1. L.M. Iyer, L. Aravind, P. Bork, K. Hofmann, A.R. Mushegian, I.B. Zhulin, and E.V. Koonin (2001) *Quoderat demonstrandum?* The mystery of experimental validation of apparently erroneous computational analyses of protein sequences, *Genome Biol*., **2**(12):research0051.
2. D.I. Gizatullina and D.G. Naumoff (2009) Reclassification of GH13 family of glycoside hydrolases. *Proceedings of the International Moscow Conference on Computational Molecular Biology*. July 20-23, 2009. Moscow. Russia. P. 249–250. (http://mccmb.belozersky.msu.ru/2009/MCCMB09_Proceedings.pdf).
3. D.G. Naumoff and M. Carreras (2009) New program PSI Protein Classifier automatizes the PSI-BLAST results analysis. *Mol. Biol.* (Engl. Transl.), **43**(4):652–664.
4. D.G. Naumoff (2006) Development of a hierarchical classification of the TIM-barrel type glycoside hydrolases. *Proceedings of the Fifth International Conference on Bioinformatics of Genome Regulation and Structure*. July 16-22, 2006. Novosibirsk. Russia. **1**:294–298. (http://www.bionet.nsc.ru/meeting/bgrs2006/BGRS_2006_V1.pdf).
5. D.G. Naumoff (2011) Hierarchical classification of glycoside hydrolases. *Biochemistry (Moscow)*, 76(6) (in press).

Computational methods for modeling AP/MS protein-protein interaction data

Alexey Nesvizhskii

University of Michigan, Ann Arbor, Michigan 48105 USA, nesvi@umich.edu

Protein-protein interaction (PPI) networks are a representation of biological systems in terms of physical interactions between functionally related proteins. There are two major types of experimental platforms for analyzing PPI networks: methods capturing direct binary interactions, as exemplified by yeast two-hybrid, and affinity purification and mass spectrometry (AP/MS) for capturing both direct and indirect interactions. Although AP/MS has been introduced relatively recently, advances in tandem mass spectrometry and efficient affinity purification strategies have already made AP/MS a commonly used technology for the analysis of PPI networks and protein complexes, including in high-throughput studies. While high-throughput AP/MS datasets are being generated with increasing popularity, computational data analysis and visualization often rely on simple approaches that do not explicitly reflect the data structure intrinsic to the experimental platform of AP/MS. Common computational challenges presented by AP/MS datasets include the need to access the statistical significance of protein interactions and filter out non-specific (background) proteins, clustering algorithms for reconstruction of protein complexes and subnetworks, and quantitative analysis of changes in the composition of protein complexes under different conditions. Here we present our recent developments in this area.

Computational methods for AP/MS protein interaction data presented here take advantage of the label-free quantitative protein abundance information such as spectral counts. This information can be easily extracted from MS data, and does not require any modification in

the experimental protocols. We will illustrate our novel methods for modeling AP/MS protein interaction data in several contexts. First, we are pursuing computational strategies for addressing the problem of false positive interactions. We will present a novel statistical approach, Significance Analysis of Interactome (SAINT) [1], which utilizes label-free quantification and Bayesian statistical modeling for assigning a confidence measure to individual interactions in both small scale and large-scale interactome studies. We will further illustrate the methods in the context of our recent work on the reconstruction of the global protein kinase and phosphatase interaction network in yeast [2].

Second, we will discuss the use of label-free quantification (spectral counts) in clustering of AP/MS protein interaction data for improved reconstruction of protein complexes. We will present a novel clustering approach [3], nested clustering, which essentially performs a two-step sequential clustering (biclustering): creation of bait clusters based on the common patterns of spectral count data across all prey proteins, and identification of nested clusters of preys sharing a similar abundance level in each one of the bait clusters. Our approach has several important advantages. Nested clustering groups the prey proteins with similar spectral counts within each bait cluster, and provides an economical expression of abundance levels by identifying a small number of discrete categories (e.g. negligible, low, medium, or high abundance). The outcome is easily interpretable in terms of the participation of each prey in one or multiple protein complexes. Second, the output is reported in the form of a bicluster, which allows individual proteins to belong to multiple complexes either as baits or as preys. Third, the method automatically chooses an optimal number of bait clusters and nested prey clusters by an extensive survey of different clustering models in terms of their likelihood - an important feature often omitted or addressed without proper statistical summaries in generic clustering algorithms. The methods will be illustrated using several human protein interaction datasets, including publicly available data from a study on chromatin remodelling complexes.

This work was been supported in part by the National Institutes of Health (R01-GM09423).

1. H. Choi et al. (2011) SAINT: probabilistic scoring of affinity purification–mass spectrometry data. *Nature Methods* **8**: 70-73.
2. A. Breitkreutz et al. (2010) Global architecture of the yeast kinome interaction network, *Science* **328**:1043-1046.
3. H. Choi et al. (2010) Analysis of protein complexes using model based biclustering of label-free quantitative AP-MS data, *Mol. Syst. Biology* **6**:385.

Prediction of neuropeptide genes in Trichoplax genome

Mikhail Nikitin[1], Leonid Moroz[2]

[1]*Lomonosov Moscow State University, A.N. Belozersky Institute of physico-chemical biology, Russia,* nikitin.fbb@gmail.com
[2]*University of Florida, Whitney Laboratory for Marine Bioscience, United States*

Background: Neuropeptides and peptide hormones are indispensable part of repertoire of regulatory ligands in animals. Prediction of these peptides in sequenced genomes is extremely difficult due to several reasons. Neuropeptides are synthesized as part of longer peptide, preprohormone, and posttranslationally cleaved by specific endopeptidases. Mature peptides are short, some only 4 aa, while non-functional parts of preprohormone are long and poorly conserved. Therefore, homology-based search generally could not predict neuropeptide genes.

Recent advances in invertebrate genomics, especially sequencing of *Nematostella* genome, demonstrated that many important vertebrate molecular tools are ancient and were present in common ancestor of bilaterians. Other invertebrate genomes, such as hemichordate *Saccoglossus* and placozoan *Trichoplax*, are also sequenced, but still poorly annotated. We choose Trichoplax to our attempt of neuropeptide gene prediction. This animal is very simple, lacking gut, nerves and muscles, but posess many developmental regulators shared with bilaterian animals, and considered to be secondarily simplified. It is interesting to find the traces of peptide neurotransmitters in its genome.

Results: We have developed and used combined search strategy, including BLAST search using known neuropeptides as query, TargetP prediction of secretory pathway signal peptide, search for regular dibasic sites in protein sequence and search for conserved domain structures of large neuropeptides. In total, we found 17 putative neuropeptides and other secreted signalling protein ligands. These include:

- one insulin family member, most similar to *C. elegans* ins-18 insulin-like peptide
- one granulin family member, with 40% amino acid identity to human granulin
- 7 precursor proteins for short peptides, three of them share limited similarity to *Aplysia* pedal peptide, PRQFVamide and FMRFamide-related peptide, other four are unique for *Trichoplax*
- 6 proteins similar to temptin, water-borne feromon modulator of molluscs, related to epidermal growth factor family
- 2 EGF (epidermal growth factor) family signalling proteins

Among these predicted ligands, granulin and EGF families are involved in embryogenesis and development, which was never observed in *Trichoplax*. Short peptides are

most intriguing, because this peptide class are directly involved in nerve system function in bilaterian animals. Immunoassays with anti-RFamide antibodies detected a number of RFamide-positive cells near the rim of animal. It is tempting to test our peptide predictions with peptide mass spectrometry, *in situ* hybridization and electrophysiological methods. It is possible to find some functional remnants of nerve system in

Temptins were earlier described only in molluscs. Six predicted temptin-related genes in Trichoplax push the origin of this protein family back to the common ancestor of Bilateria and *Trichoplax*. Only one similar protein was found outside molluscs and *Trichoplax* – in flatworm *Schistosoma*. Temptin in Aplysia serves as pheromone carrier and modulator, important for mating behavior, so we can expect role of Trichoplax temptins in collective feeding behavior or still unobserved mating.

1. K. Sonmez et al. (2009) Evolutionary sequence modeling for discovery of peptide hormones, *PLOS Computational Biology*, **5(1)**:e1000258.
2. P. Schuchert. (1993) Trichoplax adhaerens (Phylum Placozoa) has Cells that React with Antibodies Against the Neuropeptide RFamide, *Acta Zoologica*, **74:**115–117.
3. S.F. Cummins et al. (2007) Aplysia temptin - the 'glue' in the water-borne attractin pheromone complex. *FEBS Journal*, **274(20)**:5425-5437

New formulations for the genome assembly problem

Sergey Nikolenko[1], Max Alekseyev[2]

[1] *Academic University, St. Petersburg, Russia, sergey@logic.pdmi.ras.ru*
[2] *University of South Carolina, Columbia, SC, U.S.A., maxal@cs.ucsd.edu*

Traditionally the algorithmic formulation of the de novo genome assembly is posed as the Shortest Superstring Problem (SSP) that for a given set of strings (reads) asks to find a shortest string (genome) that contains them all as substrings. Both the overlap-layout-consensus and Eulerian path approaches address the genome assembly problem as an SSP [2,3]. While SSP seems to capture the genome assembly challenges, in the presence of errors its solution becomes less relevant. Moreover, users of existing genome assemblers (that all are ultimately based on the SSP) have recently begun to criticize their results. For example, in [1] the authors blame assemblers for missing repeats and, in general, producing *too short* contigs – which is, from the SSP standpoint, exactly what an assembler is supposed to do. In the current work, we propose a new mathematical and algorithmic formulations of the genome assembly problem.

We employ a Bayesian approach to specify how errors appear in the sequenced reads. Having a suitable generative model for genome sequencing, one can reverse the model to yield maximum likelihood strings (contigs). We propose a simple but very general generative model

for paired reads. Given a set of distributions (p_S, p_r, p_d, p_l, p_{err}) and a number N, we (1) generate an input genome string s, according to the distribution p_S; (2) generate N start positions $\{i_1,\ldots,i_N\}$ of reads, according to the distribution p_r; (3) generate N insert lengths $\{d_1,\ldots,d_N\}$ according to the distribution p_d and generate $2N$ read lengths $\{l_{11}, l_{12}, l_{21}, l_{22},\ldots, l_{N1}, l_{N2}\}$ according to the distribution p_l; (4) let R:={}; for j=1 to N, take a pair of substrings ($s[i_j, i_j+l_{j1}]$, $s[i_j+d_j, i_j+d_j+l_{j2}]$) and introduce errors, according to p_{err}, to obtain a pair of reads (r_{j1}, r_{j2}), and add it to R.

The resulting set R contains set of reads. It appears that the distribution p_r can be assumed uniform for most sequencing projects; p_d can be taken as a normal or logistic distribution with the parameters learned from a specific sequencer's datasets; p_l, also learned from available sequencing projects; p_{err}, approximated as a Poisson distribution whose parameters are learned from specific sequencing processes. The most interesting part of learning in this model is determination of p_S; this requires insights into the structure of a species genome and mutations distribution. The genome prior p_S is exactly the reason why a Bayesian solution is different from simply providing the shortest "reasonably correct" string– the shortest string badly fits our prior biological knowledge (which is exactly what [1] is about).

We believe that specifying a good approximation to p_S is a worthy (unsolved) problem by itself, but for the Bayesian problem of genome assembly, even a crude approximation may provide good results. As a particular approximation, we propose to use a distribution on genome *lengths* which can be inferred from any database of already assembled genomes for a certain species family. This simple prior already contributes to preferring the correct length, imposing the correct multiplicity of repeats.

We pose the Bayesian genome assembly problem: given a set of reads R and distributions (p_S, p_r, p_d, p_l, p_{err}), find the maximum likelihood genomic sequence s. It appears that for reasonably realistic distributions, the problem will not be computationally tractable. On the other hand, perhaps even a simplification (say, under the naive Bayes assumptions) may provide better results than existing SSP-based assemblers. Another application of the Bayesian approach is a new way of evaluating the quality of the assembled contigs by assessing how well the input reads fit the contigs under the probabilistic generative model.

1. C. Alkan, S. Sajjadian, and E.E. Eichler (2011) Limitations of next-generation genome sequence assembly, *Nature Methods*, **8**(1):61-65.
2. P. Pevzner (2004) *An Introduction to Bioinformatics Algorithms* (The MIT Press).
3. P. Pevzner, H. Tang, and M.S. Waterman. An Eulerian path approach to DNA fragment assembly, *Proceedings of the National Academy of Sciences*, USA, **8**(17):9748-9753, 2001.

An approach to predict cis-regulatory modules and identify conserved regulatory grammar in eukaryotic genomes

Anna NIKULOVA[1,2], Alexander FAVOROV[3,4] and Andrey MIRONOV[1,2]

[1]*Department of Bioengineering and Bioinformatics, Moscow State University, Moscow, Russia,* nikanka@gmail.com
[2]*Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia*
[3]*Department of Oncology, School of Medicine, Johns Hopkins University, Baltimore, MD, USA*
[4]*Research Institute for Genetics and Selection of Industrial Microorganisms, Genetika, Moscow, Russia*

Identification of transcriptional regulatory regions and tracing their internal organization are important for understanding of the eukariotic cell machinery. Cis-regulatory modules (CRMs) of higher eukaryotes are believed to possess a regulatory 'grammar', or fixed arrangements of binding sites, that is crucial for proper regulation and thus tends to be evolutionarily conserved [1].

Here, we present a method CORECLUST (COnservative REgulatory CLUster STructure) that predicts CRMs based on candidate sites identified using positional weight matrices. Given regulatory regions of orthologous genes CORECLUST constructs a model for the CRM by revealing the conserved rules that describe the relative location of binding sites. Then, the constructed model can be used for the prediction of CRMs with similar structure throughout the genome and for the investigation of the regulatory grammar of the system.

Application of CORECLUST to well-studied Drosophila developmental systems shows that given only one group of orthologous genes as the input, the algorithm predicts a considerable fraction of co-regulated genes. We demonstrate the advantage of accounting for the relative site arrangement, but not only site frequencies, in the CRM model and provide examples of observed conserved regulatory grammar of Drosophila early developmental enhancers, some of which were documented before.

We show that different orthologous groups of genes are characterized by distinct regulatory grammars. Also, we demonstrate the similarity of the regulatory regions of similarly expressed genes, which argues in favour of importance of the regulatory grammar for the transcriptional regulation.

Compared with related methods, CORECLUST shows similar or better performance at identification of meaningful CRMs, and it reveals internal CRM grammar.

We are grateful to Dmitri Pervouchine and Mikhail Gelfand for useful discussions and encouragement, and to Dmitry Vinogradov for technical assistance.

1. D.Papatsenko , Y.Goltsev and M.Levine (2009) Organization of developmental enhancers in the Drosophila embryo, *Nucleic Acids Research*, **37(17):**13–16.

Effect of intervention in the protection of the population of the Novosibirsk region of the influenza epidemic

Lily Nizolenko, Alexander Bachinsky
SRC VB "Vector", Koltsovo, Novosibirsk region, 630559, Russia, nizolenko@vector.nsc.ru

To simulate the influenza epidemic, a SEnImRF type model was developed. The model allows estimations of impacts of anti-epidemic measures and resources. The program of flu epidemic simulation, named "Epica" is implemented as a Windows-based application.

In this study, we have simulated the epidemic of influenza in the Novosibirsk region. The impact of the following options was estimated:

- lack of resources (health workers, opportunities for hospitalization, drug supplies);
- immunity level in the population (natural and vaccination generated);
- time of administration and the intensity of quarantine measures.

Value $\mathbf{P_I = (N_0 - N_I)/N_0\ x100}$ are used to estimate efficacy of countermeasures. Here N_0 is a value of some characteristic of the epidemic (eg the number of people infected, the number of deaths, etc.) by the end of the calculation in the absence of countermeasures, N_I is a value of the same characteristic in the implementation of countermeasure 'I'.

The lack of health professionals and places of hospitalization has no significant effect on the flu epidemic. More serious impact gives a shortage of medicines. The dependence of the level of protection of the population on the availability of drugs is shown in Fig.1

Fig. 2 compares the level of protection in different proportion of vaccinated in the population. It is shown, that 70% vaccination provides almost 100% protection (for the pathogen with average number of infected per one ill person R_0=2).

The investigation of dynamics of the epidemic with increasing intensity of measures that reduce the number of contacts (masks, quarantine) shows that the increase in this index for each 10% yields a lower total incidence of more than 100 000 cases and prevent about 2000 deaths.

Thus, comparing the protection level of the population provided by different interventions, it is possible to assess the impact of each one on the development of the epidemic (Table 1).

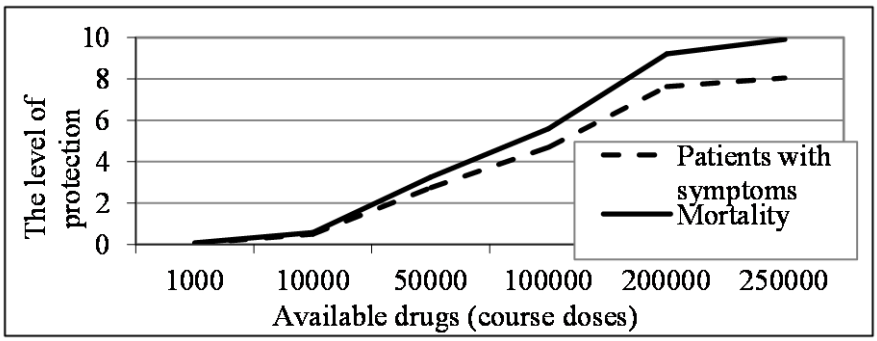


Fig.1.The dependence of the level of protection of the population on the availability of drugs.

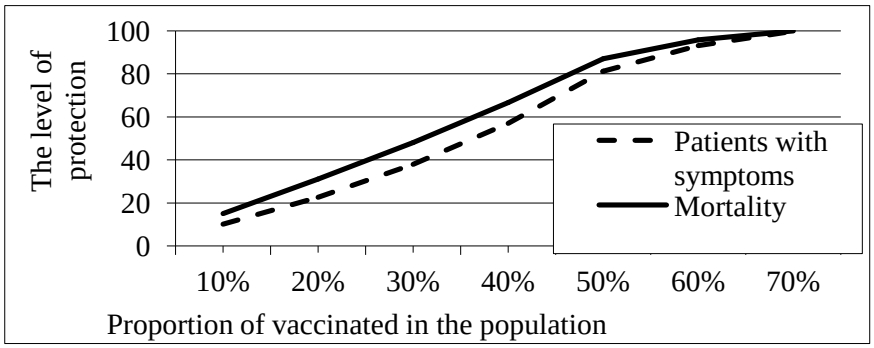


Fig. 2 The protection level according to the proportion of vaccinated.

Table 1.Comparison of the levels of protection, provided by various interventions.

Interventions	The intensity of the countermeasures	Protection in " the number of patients with symptoms"	Protection in "number of deaths"
Treatment of drugs	~400 course doses per 10 thousand population ~1000 course doses per 10 thousand population	4,7 8,04	5,6 9,9
Limiting the number of contacts	0,1 0,4	8,24 47,4	7,79 46,6
Deadline for implementing the 40% of quarantine	for the 1 day for achieving epid. threshold	50,2 47,4	49,4 46,6
Proportion of vaccinated in the population	10% 40%	10,2 54,5	15,1 64,6

Scenarios of development of epidemic of the smallpox which has arisen because of bioterrorist attack in St.-Petersburg

Lily Nizolenko, Alexander Bachinsky, Alexander Safatov
SRC VB "Vector", Koltsovo, Novosibirsk region, 630559, Russia, nizolenko@vector.nsc.ru

In Russia in general and especially cities with a high density and high populations, such as St.-Petersburg (population 6.5 million), bioterrorist attack may represent a serious danger. Currently the Hospital of S.P.Botkin is the only infectious hospital in St.-Petersburg, providing epidemiological welfare of a city. In addition there are 20 TB dispensaries. Thus, in the event of an outbreak of smallpox in the city on the most optimistic assumptions, will be available for no more than 500 - 600 places in the isolation wards of high level of protection. In this study, we investigated the feasibility of building a new hospital in order to increase this number in 2-3.

In the SRC VB "Vector" have been previously developed a model describing the effect on the dynamics of outbreaks of smallpox countermeasures such as mass vaccination and vaccination of groups at risk, isolation/observation of patients and contact persons, and based on this model, a computer program that runs on Windows.

Development and consequences of a smallpox epidemic in the presence of 500 and 1500 places in the isolation wards for various numbers of initially infected persons have been compared in simulations. The following conditions were considered:

- Collective immunity to the disease - 10%, i.e. at the level of innate immunity.
- The average number of infections from one patient $R_0 = 8$.
- Anti epidemic measures (AEM) and vaccination start when the number of hospitalized patients exceeds 400 persons, but not later than on 40^{th} day after bioterrorist attack.

Case 1 - number of initially infected persons is 10 - 100. AEM start on 30 - 40^{th} day after the bioterrorist attack (depending on the number of initially infected persons), vaccination of group at risk - on 32 - 42 day, general mass vaccination - on 34 – 44 day.

When the number of initially infected persons is low, shortage of medicines and vaccines is a critical factor. So, when there are 500 and 1500 places in the isolation wards, epidemic scenario are virtually identical. And although (if the number of primary infected persons up to 100) to the 36 day of outbreaks 500 places in probationary wards is not completely enough, it does not affect the epidemic development.

Case 2 - number of initially infected persons is 500 - 1000. AEM start on 14 - 16^{th} day after the bioterrorist attack, vaccination - on 17 - 23 day.

In this case (Fig.1), the problem of lack of places in the isolation wards is particularly acute. If there are 1500 places, the infection is terminated no later than at 53th day and the last patient recovers up to 96 days. If there are 500 places, despite the very rapid response to emergency situation and, consequently, the timely start of AEM, on the 100th day of modeling the epidemic still in full swing. At the 300-day epidemic almost over, but this time the total number of infected people is 5 988 227, that is infected almost the entire population of St. Petersburg.

Conclusion: The main factor that contributes to the effectiveness of outbreaks of smallpox countermeasures is the timely identification and isolation of patients and contacts. In this case, placing patients in conventional hospitals and clinics do not solve the problem. There must be a sufficient number of places in detention centers with a high degree of protection.

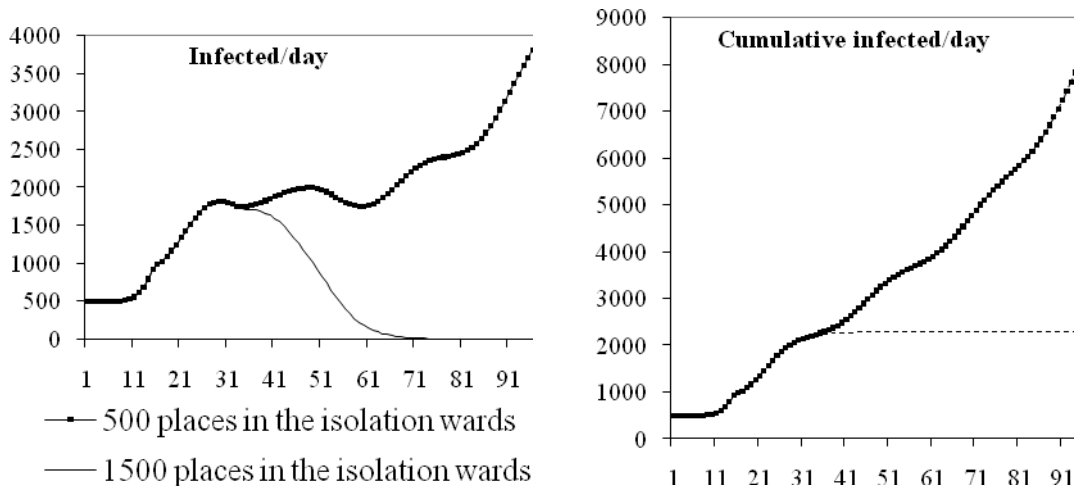


Fig 1. Current and cumulative number of infected persons per a day up to 100 days of the outbreak of smallpox, when the number of initially infected persons is 500.

Formation and functional conservation of regulatory binding site complexes in eukaryotes

Armita Nourmohammad, Michael Laessig

Institute for Theoretical Physics, University of Cologne, Germany, armita@thp.uni-koeln.de

In eukaryotes, genes often have complex regulatory input, which is encoded by multiple transcription factor binding sites linked to a common function. In this talk, we discuss the evolution of such binding site complexes. We present two case studies addressing the following questions: What mode of sequence evolution is relevant for the formation of these site complexes? How is the function of site complexes maintained during evolution? (1) In *Drosophila*, we show that local sequence duplications are a pervasive formation mode of site complexes: the majority of neighboring binding sites share a common sequence ancestor. This result is obtained by sequence analysis based on an evolutionary model with two distinct modes of binding site formation: evolution from independent sequence origins and divergent evolution following duplication of a common ancestor sequence. (2) In *S. cerevisiae*, we study the evolution of binding site complexes consisting of a strong site and neighboring weak sites. We develop a thermodynamic framework to characterize the effective affinity of site complexes to multiple transcription factors with cooperative binding. We show that there is wide-spread compensatory evolution within these complexes: individual sites are under weak selection and have a high turnover rate, whereas the overall binding affinity of the site complex is under stronger stabilizing selection maintaining its functional output.

Selection of optimal parameters for molecular dynamics computation: generating of NMR-comparable trajectories

Alex Nyporko, Aliona Yaremchuk

Taras Shevchenko Kiev National University, High Technology Institute, dfnalex@gmail.com

Molecular dynamics computation is very useful method for investigation of protein structure [1]. This approach gives a possibility to discriminate and evaluate the nuances of protein structure and behavior which are indiscernible for other methods (in particular, distant correlative motions [2] and distant effects of amino acid replacements [3]) as well as allow to observe perturbations of protein structure for some of time. However, despite usability and wide facilities, results obtained via molecular dynamics sometimes are ambiguous. Main claim to molecular dynamics is the slow but stable gradual increase of protein oscillation amplitude and,

as a result, lack of relaxation of molecular system that is observed in some cases. Thus, the aim of this research is selection of optimal criteria for molecular dynamics calculations (force fields, calculation parameters, etc.) to generate time-stable protein motions trajectories which correlate with NMR-derived motions' patterns.

Protein Data Bank was scanned to retrieve the spatial structures of water-soluble proteins, participating in cytoskeleton functioning, which are resolved by NMR spectroscopy and have no less than 20 deposed conformers. Among them the four proteins were selected in a random way for 50 ns molecular dynamics calculation with various criteria/conditions (PDB access codes are 1UNC, 1AJ3, 1GM1 and 1T0Y). Calculations were carried out with GROMACS 4.0.3 software using force fields Gromos 96, Gromacs and OPLS. The next parameters were varied: *nstlist* – frequency to update the neighbor list (0, 1, 10, 100), *comm_mode* – center of mass motion removal function (none, linear, angular), *T-coupling* – thermostat function (berendsen, V-rescale), *rcoulomb* – computation method for electrostatic interactions (cut-off, PME). Structural stability was estimated by conformational energy dynamics (using *g_energy* module), and levels of molecular oscillations (using *g_rms* module). Motion patterns for each studied protein were obtained from 20 molecular dynamics' conformers with lowest potential energies and compared with appropriated patterns calculated from NMR-derived conformers.

Among studied parameters of molecular dynamics calculations PME method of electrostatic interactions evaluations shown the most stabilize effect on protein dynamics. Application of this method in combination with any others resulted in stable horizontal plateau of molecular oscillations and decrease of average oscillation level and oscillation amplitude. Using V-rescale thermostat resulted in accelerated stabilization of motions level. Variation of *nstlist* hasn't produced significant consequences (except *nstlist*=1 resulted in accelerated stabilization of system similar to V-rescale action). Effects of center of mass motion removal function were ambiguous. Among studied force fields OPLS appeared to be the most optimal for molecular dynamics calculations. Profiles of oscillations of individual amino acid residues obtained from trajectories calculated with OPLS are similar to NMR motion patterns for all studied protein. Using Gromacs force field revealed good correlation with NMR data for 1GM1 and 1T0Y proteins, using Gromos 96 force field – with NMR data for 1T0Y only. Thus, application of OPLS force field with PME electrostatic calculation and V-rescale thermostat seems to be optimal for molecular dynamics calculation and gives a possibility to obtain the protein dynamics trajectories comparable by a quality with NMR data.

1. M. Meli, G. Colombo (2009) Molecular simulations of peptides: a useful tool for the development of new drugs and for the study of molecular recognition, *Methods Mol Biol*, **570**:77-153.
2. O. F. Lange, H. Grubmaller (2008) Full correlation analysis of conformational protein dynamics, *Proteins*,**70***:*1294-312.
3. P. Lazar et al. (2009) Molecular modeling study on the effect of residues distant from the nucleotide-binding portion on RNA binding in *Staphylococcus aureus* Hfq, *J Mol Graph Model.* **28**:253-260.

Housekeeping genes in the human genome: what about cancer?

Roman Tychko, Anna Kudryavtseva, Nina Oparina

Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, 32 Vavilov street, Moscow, Russia, mashkova@eimb.ru, oparina@gmail.com

The term “housekeeping genes” is widely used in modern biology. Mostly such gene are proposed to be involved in cellular processes common for all cells despite their tissue differentiation. These genes are probably active during all stages of ontogenesis. Well-known “housekeeping” processes include glycolysis, aminoacides and nucleotides biosynthesis, protein catabolism etc. Candidate housekeeping genes are mostly predicted basing on comparative transcriptomic analysis in a set of adult organism normal tissues. The available dataset consists mostly of microarray data and ESTs, both methods are suitable mostly for highly expressed mRNAs. Appropriate data could be obtained using RNAseq in the nearest future. We have analysed current sets of human candidate housekeeping genes (selected basing on microarray data of normal tissues) in attempt to detect features associated with the stability of mRNA level of the gene in cancer cells. Several (~15) housekeeping genes are widely used as “reference genes” for normalization of gene expression. It is known form biomedical publications that no one from these genes is really “stable” in all pairs of studied normal and cancer tissues. For example, GAPDH gene expression is disregulated in such cancer types like colorectal cancer characterized with activated glycolysis. Are there any really “stable” housekeeping genes probably useful for most of known neoplasms? We have carried out comparative analysis of dataset of microarrays and selected ~1500 genes similarly expressed in all normal tissues of adult human as well as in stem cells and embryo. Many of these genes were characterized by similar features (such as CpG-island structure, TF sites etc.) in comparison to the whole genome. Nevertheless stability of mRNA level of selected genes in five mostly studied cancers varied: mRNAs levels of only ~400 of “housekeeping genes” were both in cancers and normal tissues. We have performed comparative analysis of expression of “housekeeping genes” in lung cancer, breast cancer, colorectal cancer, prostate cancer and kidney cancer. Sequence analysis of the corresponding transcripts and genomic region demonstrated frequent aberrations of “unstable” housekeeping genes in cancers. The dataset of housekeeping genes characterized with the most stable expression in all cells (normal, cancer and embryonal) consists of genes of various functions while there structural features are similar (including intron length, CpG-island

structure etc.). Our approach is suitable for expression-free selection of candidate reference genes suitable for detection of new stable genes, including those with low mRNA expression.

1. Sullivan-Gunn M, Hinch E, Vaughan V, Lewandowski P. (2011) Choosing a stable housekeeping gene and protein is essential in generating valid gene and protein expression results. Br J Cancer. 104(6):1055
2. Wan Q, Whang I, Choi CY, Lee JS, Lee J. (2011) Validation of housekeeping genes as internal controls for studying biomarkers of endocrine-disrupting chemicals in disk abalone by real-time PCR. Comp Biochem Physiol C Toxicol Pharmacol. 153(3):259-68.
3. Wierstra I. (2008) Sp1: emerging roles--beyond constitutive activation of TATA-less housekeeping genes. Biochem Biophys Res Commun. 372(1):1-13.
4. Tu Z, Wang L, Xu M, Zhou X, Chen T, Sun F. (2006) Further understanding human disease genes by comparing with housekeeping genes and other genes. BMC Genomics. 7:31.

New reference gene for various human cancers gene expression normalization: RPN1 testing on lung and kidney cancers

Georgy KRASNOV[1], Nina J. OPARINA[1], Alexey DMITRIEV[1], Anna KUDRYAVTSEVA[1], Ekaterina ANEDCHENKO[1], Tatyana KONDRATIEVA[2], Eugene ZABAROVSKY[3], Vera SENCHENKO[1]

[1] *Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow, Russia, oparina@gmail.com*
[2] *Blokhin Scientific Center of Oncology, Russian Academy of Medical Sciences, Moscow, Russia*
[3] *Karolinska Institutet, Stokholm, Sweden*

Gene expression studies are impossible without proper normalization procedures. The most useful way is the normalization basing on the “reference” genes: such genes, mostly housekeeping, are thought to be stably expressed in most normal tissues as well as in cancers. Nevertheless, there are no universal reference genes appropriate for studying all cancer types. Adequate gene expression quantification while investigating mRNA levels of cancer-associated genes is impossible without two or more reference gene. Some of the most widely applied human reference genes include *ACTB, GAPDH, GUSB, B2M, HPRT1, TBP* and some other. By the way each of these genes is suitable only for some cancers: stable in kidney cancerogenesis *GUSB* is frequently downregulated in lung cancer, while *GAPDH* expression changes quite a contrary. We have proposed an approach based on the EST sequences analysis for prioritization of adequate candidates of new reference genes. Our rules include estimation of mRNA level stability in normal and cancer cells, frequency of splicing aberrations and mutations. The resulting scoring-function was used to select putative new reference genes for lung and kidney cancers. Further filter was based on microarray analysis (ONCOMINE) and functional annotation of selected candidate genes. 14 EST-based best scored reference genes were selected

and after filtering the *RPN1* was found as the most appropriate reference gene for both cancers. Realtime PCR studies for *RPN1* expression in non-small cell lung cancer and clear cell renal cancer were carried out. The stability of its expression was similar to *GAPDH* in lung cancer and *GUSB* in kidney cancer. Our results show that bioinformatics procedures could give us proper genes for further experimental studies.

Integrative analysis of transcription factors binding profiles regulating embryonic stem cell identity based on ChIP-seq and expression arrays technologies

Yuriy Orlov[1], Nikolay Podkolodny[1], Huck-Hui Ng[2]

[1]*Institute of Cytology and Genetics, Novosibirsk, Russian Federation, orlov@bionet.nsc.ru*
[2]*Genome Institute of Singapore, Singapore*

Human ESCs (hESCs) have the capacity for extensive self-renewal under in vitro culture conditions. A second hallmark of these cells is their ability to undergo multi-lineage differentiation, also defined as pluripotency. The robust self-renewal capability of these pluripotent cells has great importance for therapeutic applications and drug discovery (Chia et al., 2010). To investigate transcription factors responsible for stemness we used whole genome binding profiles obtained by ChIP-seq experiments in mouse and human. Previously we used 13 TF binding sets for mouse ESC (Chen et al., 2008) to reveal co-localization of key transcription factor binding sites in genome scale.

Despite the significant differences between pluripotent mammalian stem cells, the same set of transcription factors (Oct4, Sox2, Klf4 and c-Myc) can be used to reprogram human and mouse somatic cells into induced pluripotent stem cells. POU5F1 (coding for the protein OCT4) and NANOG, both key components of the core transcriptional regulatory network are highly expressed in undifferentiated ESCs and upon differentiation the expression of these genes are reduced. These and other transcription regulators, including the co-activator p300, show extensive co-localization in mouse genome.

We used set of bioinformatics methods to analyse network properties of reveled genes and study binding profiles of key transcription factors responsible for ESC maintenance in mouse and human. Recently, a genome-wide RNA interference (RNAi) screen (multiple knockdown experiments) to identify genes which regulate the self-renewal and pluripotency

properties of hESCs was reported (Chia et al., 2010). Among known transcription factors identified in the hESC screen, functional role of PRDM14 was shown. Genome-wide location profiling experiments in human revealed that PRDM14 co-localized extensively with other key transcription factors such as OCT4, NANOG and SOX2. In addition, PRDM14 can repress transcription through the recruitment of polycomb group proteins.

We used web-resources Reactome (www.reactome.org) and STRING (http://string-db.org/) for pathway analysis to reveal groups of interacting proteins among top 500 genes revealed by the screen. We found several protein complexes in the set of genes revealed by siRAN screen: INO80 complex, mediator complex, TAF complex, COP9 signalosome, eukaryotic initiation factor complex and spliceosome complex. Enriched sequence motifs were identified by de novo motif discovery programs Weeder, MEME and CisFinder, as well as custom-made scripts. All programs identified a motif overrepresented motif in PRDM14 ChIP-seq peak regions with the core 9-mer GGTCTCTAA as the most or second most enriched. Co-occurrence analysis to study the overlap of PRDM14 with other transcription factors binding sites was performed as described previously (Chen et al., 2008). CTCF, OCT4 and NANOG ChIP-seq datasets were generated and processed in the same way as the PRDM14 dataset. KLF4, MYC, p300, SOX2 and histone modifications ChIP-seq data were obtained from GEO NCBI. PRDM14 sites (ChIP-seq peaks) associated with 996 PRDM14 regulated genes (in 20Kb to the gene borders) were analyzed for co-occurring PWMs found in the TRANSFAC database. We found enrichment of OCT4, AP2 and SP1 binding proximal to PRDM14 sites.

The work was supported in part by Interdisciplinary Integration projects SB RAS 26 and 119, RFBR 11-04-01888.

1. X. Chen et al. (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Cell 133, 1106-17.
2. N.-Y. Chia et al. (2010) A genome-wide RNAi screen identifies PRDM14 as a regulator of POU5F1 and human embryonic stem cell identity. Nature, 468(7321): 316-20.

Promoter Regions of the Genes Encoding Human Macrophageal Cytokines Possess Dioxon Response Elements

E. Oshchepkova[1], D. Oshchepkov[1], E. Kashina[1], E. Antontseva [1], D. Furman[1,2], V. Mordvinov[1]

[1]*Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia*
[2]*Novosibirsk State University, Novosibirsk, Russia*
nzhenia@bionet.nsc.ru, diman@bionet.nsc.ru

TCDD (2,3,7,8-tetrachlorodibenzo-p-dioxin) – is most toxic congener of dioxin-like compounds. TCDD induces a broad spectrum of biological consequences, including disruption of normal hormone signaling pathways, reproductive and developmental defects, liver damage, wasting syndrome, and cancer [1]. Immune system is also the target for TCDD toxicity and it's well recognized that disruption of immune response due to TCCD immunotoxic influence can lead to increased incidence of some types of cancer, allergic and autoimmune diseases, etc. On the cell level, the dioxin mediates gene expression via AhR/ARNT transcription complex activation, which binds to dioxin responsive elements (DRE) in the regulatory regions of the inducible genes. Previously the TCDD–mediated modulation of the expression of genes encoding pro-inflammatory cytokines was shown experimentally [2], showing the most plausible mechanism for inappropriate enhancement of immune function. For better understanding whether AhR/ARNT action is direct, or indirect through the immanent transcription factors, we have performed the search of the putative DREs in the regulatory regions of the genes, encoding pro-inflammatory as well as anti-inflammatory cytokines, expressed in activated macrophages. SITECON software package was implemented for DREs searching [3]. DRE sites have been detected in number of promoters of macrophage pro-inflammatory cytokine genes: *IL12a, IL12b, IL15, IL24, IFNA, CCL22, MIP2A, IP10,* etc. Also, DRE sites were found in the promoter of anti-inflammatory cytokine gene *IL-4*. A few DREs' functional activity was proven with EMSA and Rt-PCR experiments.

Obtained results sound in favor of the possibility that TCDD can directly mediate the gene expression of the genes encoding macrophage cytokines, containing DREs in their regulatory regions, thus affecting the immune response. Also simultaneous TCDD–mediated direct activation of anti-inflammatory cytokine gene *IL-4* detected indicates the possibility of Th1/Th2 cytokine balance shift leading to development of wide range of pathologies, including asthma, allergic and autoimmune diseases [4].

The work was supported by the Russian Academy of Sciences (program A.II.6); Siberian Branch of the Russian Academy of Sciences (interdisciplinary integration project no. 119 and Lavrentiev's project №6.3); Federal Agency for Science and Innovations (state contract no. 16.512.11.2129); and the RFBR grant no. 09-04-12209-ofi_m.

1. P.K. Mandal. (2005) Dioxin: a review of its environmental effects and its aryl hydrocarbon receptor biology. *J Comp Physiol B.* **175**: 221-230.
2. C.F. Vogel et al. (2007) Modulation of the chemokines KC and MCP-1 by 2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD) in mice. *Arch Biochem Biophys.* **461**:169-175.
3. D.Yu. Oshchepkov et al. (2004) SITECON: a tool for detecting conservative conformational and physicochemical properties in transcription factor binding site alignments and for site recognition. *Nucl. Acids Res.* **32**: W208–212.
4. P.L. Ngoc et al. (2005) Cytokines, allergy, and asthma. *Curr Opin Allergy Clin Immunol*. **5(2):**161-6.

The Novel Approach to the Cytochrome c tertiary structure design

Tatyana Ostroverkhova[1], Rita Chertkova[2], Alexei Nekrasov[2]

[1]*Moscow State University, Leninskie gory, 1/12, Moscow, Russian Federation, tato-tato@list.ru*
[2]*Institute of Bioorganic Chemistry of the Russian Academy of Sciences, Miklukho-Maklaya, 16/10, Moscow, Russian Federation*

Cytochrome c is heme-containing protein and an essential component of the mitochondrial electron transport chain. The studing of cytochrome c tertiary structure using ANIS-method allowed to focus on site (76-83) of its amino-acid sequence [1-3]. Probably, this site of cytochrome c is of major importance for protein functioning, for instance, at respiratory chain.

The informational structure of cytochrome c resulted using ANIS-method [1-3]. According ANIS data cytochrome c has two high-ranking elements of informational structure (ELIS). Each ELIS contains one ligand interacting with iron atom of heme. ANIS-method reveals sites with different degree of informational coordination. Interestingly, informational structure of cytochrome c includes only one abnormal distribution density site (76-83) with anomalously low density of first rank ELIS (ADD- site). It is important to notice the site (76-83) carries iron ligand Met-80. On the assumption of protein physics conformational changes are located at more flexible site with great amount of free degrees. We proposed significant role of ADD- site at specific interactions with ubiquinol:cytochrome c reductase (complex III) and cytochrome c oxidase (complex IV). For verification our hypothesis we design series forms of horse heart cytochrome c with specific substitutions of ADD- site located amino acid residues.

Modeling of informational structure of cytochrome c was realized regarding part of its mutant variants. Such forms carry about site (76-83) with anomalously high density of first rank ELIS (ADD+ site) instead of ADD- site. Data from analysis recurrence of amino acid at E.coli proteome and nonhomologous proteins sites were applicated to determination amino acid substitutions. Proposed amino acid substituions change structure characteristics and provide formation more hard site (76-83) of cytochrome c. The aim of this design was to engineer the mutated forms of cytochrome c that the reactivity towards the respiratory chain components would be lost.

Computative modeling on a basis of ANIS-method allowed to suggest series of mutant variants of cytochrome c [1-3]. Mutant genes of cytochrome c I81Y/A83Y/G84N, T78N/K79Y/M80I/I81M/F82N and T78S/K79P were constructed and obtained by site-directed mutagenesis as recommended using the QuikChangeTM Mutagenesis Kit (Stratagene, USA). Recombinant proteins have been purified and their biological activity were tested at rat liver mitopalasts. Maximal decrease of succinate: cytochrome c-reductase activity occured in case of T78S/K79P mutant variant and maximal decrease of cytochrome c-oxidase activity - in case of I81Y/A83Y/G84N mutant variant. Thus, amino acid substituions were modeled by ANIS-method modified biological activity of cytochrome c. Investigation results of upper mutant variants list showed achievement for lowering cytochrome c interaction with respiratory chain and indicate to (76-83) region significance for electron-transport activity of protein. Informational structures of next cytochrome c mutated variants series were made and analyzed by ANIS-method. A few forms had pronounced ADD+ site at the (76-83) region of cytochrome c amino acid sequence. Mutant genes P76R/G77I/T78L/K79V/I81V/F82S/A83R, P76R/G77R/T78E/I81S/F82A/A83I and P76L/G77A/T78V/K79G/I81G/F82G/A83G were constructed and synthesized. At presence, search is for optimization of expressed system conditions. Recombinant proteins are needed for testing its biochemical properties and determination of designed amino acid substitutions role.

1. Nekrasov A.N. (2004) Analysis of the information structure of protein sequences: a new method for analyzing the domain organization of protein, J. Biomol. Struct. Dyn., 21(5): 615-624.
2. Nekrasov A.N. et al. (2009) Design of a novel interleukin-13 antagonist from analysis of informational structure, Biochemistry (Mosc), 74(4): 399-405.
3. A. Nekrasov et al. (2010) Structural features of the interfaces in enzyme-inhibitor complexes, J. Biomol. Struct. Dyn., 28 (1): 85-90.

Electrostatic properties of the natural genome DNA and its elements

Alexander Osypov, Svetlana Kamzolova
Institute of Cell Biophysics of RAS, Russian Federation, aosypov@gmail.com

Physical properties of the genome DNA influence its biological functions, but not so much is known about it, and the sequence textual analysis rules the (bioinformatician) world due to the progress in the techniques of sequencing, analytical algorithms and annotation. Though the text analysis on its own struggles to solve such problems as predicting and describing promoters and their strength, as well as some other genome elements. On the other hand, studies of the physical properties sometimes give valuable insights into the intimate mechanisms of the genome functioning. So due is the combination of the textual methods sensu stricto with physical properties analysis.

One of the most important physical properties of the DNA is electrostatics. Using the original calculation method, we have set up the DEPPDB database with the electrostatic properties of all the sequenced procariotic genomes with their full biological annotation, grouped by the taxonomic tree to support the comparative and evolutionary studies.

Electrostatic properties are dependent on the sequence content, but not exactly one-to-one, as different sequences may exhibit the same characteristics and vise verse. Also it dramatically depends on the flanking regions of a great length. Observations were made on the regulation of promoter function, which is dependent on the electrostatic characteristics of the core promoter and far upstream regions. Revealed is the role of electrostatic properties in the transcription factors-DNA interactions. Many other genome elements, such as terminators, show electrostatic peculiarities, provoking idea of the evolutionary universal role of electrostatics in the genome functioning. Most intriguing is the pattern around gene starts, correlating with the taxonomic proximity.

The authors are grateful to Saveljeva E. G. for technical support and the Institute of Mathematical Problems of Biology of the Russian Academy of Sciences for hosting the Database.

1. R. V. Polozov, T. R. Dzhelyadin, A. A. Sorokin, N. N. Ivanova, V. S. Sivozhelezov, S. G. Kamzolova (1999) , J. Biomol. Struct. Dyn. 16(6):1135-1143.
2. S.G. Kamzolova, A.A. Sorokin, T.D. Dzhelyadin P.M., Beskaravainy, A.A. Osypov (2005) , J. Biomol. Struct. Dyn. 23(3):341-346.
3. S. G. Kamzolova, V. S. Sivozhelezov, A. A. Sorokin, T. R Dzhelyadin, N. N. Ivanova, R. V. Polozov (2000) , J. Biomol. Struct. Dyn. 18(3):325-334.
4. A. A. Sorokin, A. A. Osypov, T. R. Dzhelyadin, P. M. Beskaravainy, S. G. Kamzolova (2006) , J. Bioinform. Comput. Biol. 4(2):455-467.
5. A. A. Osypov, G. G. Krutinin, S. G. Kamzolova (2010) , J. Bioinform. Comput. Biol. 8(3): 413-425.

New insights into protein-DNA electrostatic interactions: beyond promoters to transcription factors binding sites

Eugenia Krutinina, Gleb Krutinin, Svetlana Kamzolova, Alexander Osypov
Institute of Cell Biophysics of RAS, Russian Federation, krutininaea@gmail.com

Electrostatic properties of genome DNA influence the primary recognition and regulation of transcription by RNA-polymerase. This enzyme may identify promoters and evaluate their strength due to the peculiarity of their electrostatic profiles. To reveal the role of electrostatic properties also in the transcription factors-DNA interactions we studied binding sites of different families of these proteins.

The analysis of the profiles using DEPPDB Database showed some common features, illustrated here with the CRP binding sites in the E.coli genome DNA.

The averaged profiles of the DNA electrostatic potential aligned around the CRP binding sites centers exhibit the pronounced rise in the negative potential value with the characteristic profile in the consensus area of 16 bp. The extensive ~300 bp long symmetrical overall potential rise can't be explained by the influence of the consensus itself and reflects the sequence organization of the flanking regions, contributing to the high potential area formation. Apparently that was selected evolutionary to support the binding site recognition by the regulation protein molecule and its retention.

The same overall properties, though vary in particular details, are typical to binding sites of other families of transcription factors in a diverged range of bacterial taxa.

These data reveal the role of electrostatic properties of DNA in the recognition of the transcription regulation proteins binding sites, further confirming their universal importance in the protein-DNA interactions beyond the classical promoter-RNA polymerase recognition and regulation, validating the studies of the electrostatic properties of genome DNA in addition to the traditional textual analysis of its sequence.

The authors are grateful to Saveljeva E. G. for technical support and the Institute of Mathematical Problems of Biology of the Russian Academy of Sciences for hosting the Database.

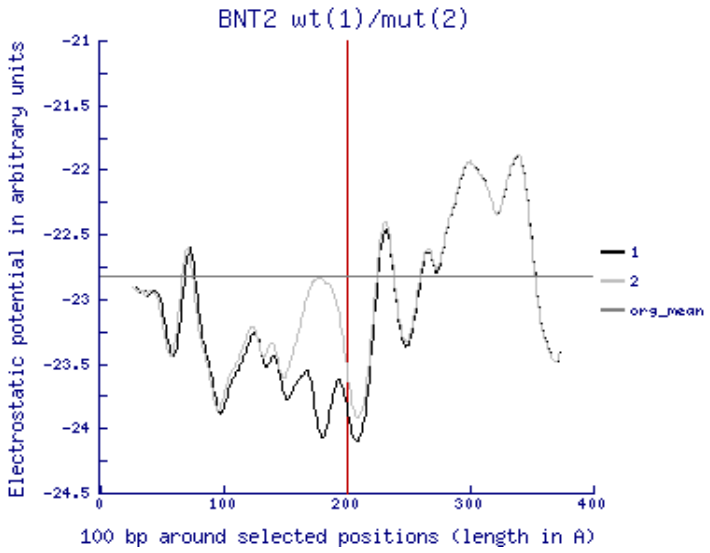
1. R. V. Polozov, T. R. Dzhelyadin, A. A. Sorokin, N. N. Ivanova, V. S. Sivozhelezov, S. G. Kamzolova (1999) , J. Biomol. Struct. Dyn. 16(6):1135-1143.
2. S.G. Kamzolova, A.A. Sorokin, T.D. Dzhelyadin P.M., Beskaravainy, A.A. Osypov (2005) , J. Biomol. Struct. Dyn. 23(3):341-346.
3. S. G. Kamzolova, V. S. Sivozhelezov, A. A. Sorokin, T. R Dzhelyadin, N. N. Ivanova, R. V. Polozov (2000) , J. Biomol. Struct. Dyn. 18(3):325-334.
4. A. A. Sorokin, A. A. Osypov, T. R. Dzhelyadin, P. M. Beskaravainy, S. G. Kamzolova (2006) , J. Bioinform. Comput. Biol. 4(2):455-467.
5. A. A. Osypov, G. G. Krutinin, S. G. Kamzolova (2010) , J. Bioinform. Comput. Biol. 8(3): 413-425.

Electrostatic properties complement the DNA bending in the Escherichia coli O157:H7 pO157 plasmid BNT2 promoter functioning

Eugenia Krutinina, Alexander Osypov
Institute of Cell Biophysics of RAS, 142290, Pushchino, Moscow Region, Russia, aosypov@gmail.com

DNA physical properties are well known to influence its biological functions. It was shown that electrostatic properties of genome DNA influence the primary recognition and regulation of transcription by RNA-polymerase. This enzyme may identify promoters and evaluate their strength due to the peculiarity of their electrostatic profiles.

Another studied parameter that influences promoter strength is the DNA bending in the promoter area. Yoon et al. (2010) showed that the unusual *Escherichia coli* O157:H7 pO157 plasmid BNT2 promoter relies on the DNA bending in its functioning. This promoter has AAAAA tract in the spacer and an unusual AAAAAT -10 element. The authors substituted the spacer AAA tract with the TCG sequence and found that the resultant promoter a) posses lesser bending and b) has much lower activity. As the activity of the native bent promoter strongly reduces with the temperature rise above the curvature-relaxing point the authors make conclusion that the promoter activity relies on the intrinsically bent DNA structure of the AT tracts. They confirmed this by a) calculating the curvature by an available program and b) by studying of the DNA retardation in the electrophoresis under different temperatures. However, the point was missed concerning the twice lower activity of the mutant promoter under the temperature above the curvature-relaxing point, which can not be accounted for the bending.



We have calculated the electrostatic potential profile along the native and mutant sequences in this promoter area. The native promoter with AT tract posses a prominent rise in the electronegative potential value along it, that correlates with the promoter strength and probably facilitates the promoter recognition and binding. In the mutant with the part of the A tract substituted with

the TCG sequence there is no such a rise in this place (see figure), that obviously affects the DNA-RNA polymerase interactions.

Thus the electrostatic properties of the promoter DNA complement the DNA bending in the *Escherichia coli* O157:H7 pO157 plasmid BNT2 promoter functioning.

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

1. R. V. Polozov, T. R. Dzhelyadin, A. A. Sorokin, N. N. Ivanova, V. S. Sivozhelezov, S. G. Kamzolova (1999), *J. Biomol. Struct. Dyn.* **16(6):**1135-1143.
2. S.G. Kamzolova, A.A. Sorokin, T.D. Dzhelyadin P.M., Beskaravainy, A.A. Osypov (2005), *J. Biomol. Struct. Dyn.* **23(3):**341-346.
3. S. G. Kamzolova, V. S. Sivozhelezov, A. A. Sorokin, T. R Dzhelyadin, N. N. Ivanova, R. V. Polozov (2000), *J. Biomol. Struct. Dyn.* **18(3):**325-334.
4. A. A. Sorokin, A. A. Osypov, T. R. Dzhelyadin, P. M. Beskaravainy, S. G. Kamzolova (2006), *J. Bioinform. Comput. Biol.* **4(2):**455-467.
5. A. A. Osypov, G. G. Krutinin, S. G. Kamzolova (2010), *J. Bioinform. Comput. Biol.* **8(3):** 413-425.
6. J. W. Yoon, M. K. Park, C. J. Hovde, Seung-Hak Cho, Jong-Chul Kim, Mi-Sun Park, W. Kim (2010), *Biochem. Biophys. Res. Commun.* **391(4)**:1792-1797.

In silico analysis of the interaction of new nitro- and dinitroaniline compounds with oat α-tubulin

Sergey Ozheredov, Pavel Karpov, Oleg Demchuk, Alla Yemets, Yaroslav Blume

[1]*Institute of Food Biotechnology and Genomics, Natl. Acad. Sci. of Ukraine, Ukraine, ozheredov@gmail.com*

An increasing interest to antimicrotubular compounds such as the dinitroanilines and phosphorothioamidates is closely associated with their herbicide and antiprotozoan activity [1]. Earlier we have performed *in silico* design and synthesis of nitro- and dinitroaniline compounds [1, 2]. The main goal of current *in silico* study was a screening of these new biologically active nitro- and dinitroaniline derivatives through assessment of the stability of their complexes formed with plant α-tubulin.

The α-tubulin from oat (*Avena sativa*) (UniProt: Q38771) was used as molecular target. Using homologous modeling in Swiss-PdbViewer v4.0.1. (http://www.expasy.org/spdbv/), a three-dimensional model of α-tubulin were constructed. The αβ-tubulin heterodimer (PDB: 1TUB) from *Sus scrofa* was chosen as template structure. Accordingly to the dinitroaniline binding site data [3], the protein-ligand complexes were designed with Accelris DS Visualizer 2.0 (www.accelrys.com) package. Molecular dynamics simulation in GROMACS and it's estimation was performed according to root-mean-square deviation (RMSD) of atom

distances, absolute binding free energy (ΔG_{bind} - the free energy difference between the bound and unbound states) [4], and the average number of hydrogen bonds between ligand and receptor [5].

The data of ligand conformational energy changes in water and binding site reflect the affinity to receptor [5, 6]. According to molecular dynamic results, the ΔG_{bind} values were ranged between -47.80 to -315.14 kJ/mol. Therefore, these ΔG_{bind} values confirmed probability of α-tubulin complexes formation. It was established that 4-methylsulfonyl-2,6-dinitroaniline, N'-(N''-[2,6-dinitro-4-trifluoromethylphenyl]propyl) morpholine (Br-44), N*1*-(2,6-dinitro-4-trifluoromethyl-phenyl)-ethylene-1,2-diaminhydrochloride (CNA-017) and 1-[2-(2,6-dinitro-4-trifluoromethyl-phenylamino)-ethyl]-3-ethylthiocarbamide (CNA-030) are the most potential biologically active compounds.

The mean value of ΔG_{bind} was in the range from -239,25 to -315,14 kJ/mol, and significantly exceed these rate in comparison with trifluralin (control) – -113,37 kJ/mol. Accordingly to this observation, new compounds posses by high antimitotic activity. Based on the number of hydrogen bonds in the ligand-protein interactions, derivatives of nitrobenzene, as 3-(4-ethoxy-2-nitro-phenylcarbomoil)-acrylic acid (CNA-004), 3-(4-methyl-2-nitro-phenylcarbomoyl)-acrylic acid (CNA-005) and 3-(4-amino -3-nitro-phenylcarbomoil)-acrylic acid (CNA-006) were most effective. Average number of hydrogen bonds formed by these compounds is 4-5, with mean value of ΔG_{bind} amounted to -171.06, -171.15 and -179.80 kJ/mol, respectively. We can conclude that selected compounds (CNA-004, CNA-005, CNA-006, CAN-017, CAN-030) are most perspective for further testing of their antimicrotubular activity.

1. B.M. Britsun et al. (2009) 2,6-Dinitroanilines: synthesis, herbicidal and antiprotozoan properties, Ukr. Bioorg. Acta, 7(1):16-27 (in Ukrainian).
2. S.P. Ozheredov et al. (2009) Screening of new 2,4- and 2,6-dinitroaniline derivates for phytotoxicity and antimitotic activity, Cytol. Genet., 43(5):3-13.
3. A.Yu. Nyporko et. al. (2009) Structural-biological characteristics of tubulin interaction with dinitoanilines, Cytol. Genetics, 43(4):56-70.
4. M. Almlof et al. (2004) Binding affinities prediction with different force fields: examination of the linear interaction energy method, J. Comput. Chem., 25(10):1242-1254.
5. J. Gu, P. Bourne (2009) Structural bioinformatics, 2nd Ed., 1035p. (John Wiley & Sons.).
6. K. Ramachandran et al. (2008) Computational chemistry and molecular modeling principles and applications, 387p. (Springer-Verlag: Berlin).

Computer-based search for promoters within the A/T-rich genome of *Helicobacter pylori*

S.S. Kiselev, O.N. Ozoline

Institute of Cell Biophysics, Russian Academy of Sciences, Pushchino, Russia, ozoline@icb.psn.ru

An ability of promoter finders to identify all transcribed regions in the genome, regardless of what type of products (mRNAs, rRNA, tRNA or various untranslated RNAs) are encoded, makes them an important tool for annotation. However a requirement for certain adaptation of a computer programs to the consensus elements of particular promoters hampers their utilization for genomes with poorly studied regulatory regions. At the same time, evolution stability of the transcription machinery, imposing certain restrictions on conformational features of the promoter sites, provides a chance to use structure-specific modules as the main indicators of promoter DNA. The first version of unified software has been suggested [1] on the basis of promoter finder PlatProm, initially adapted for σ^D-dependent promoters of *E.coli*. Position weight matrices (PWM) of PlatProm score both sequence-specific elements, recognized by σ^D, and structure-specific modules, favoring transcription complex formation [2]. In unified version (PlatPromU) of the program σ^D-specific PWMs were switched off. This version accurately identified 79.7% promoters in the genome of evolutionarily distant from *E.coli* bacterium *C.glutamicum* with $p<0.001$ reliability, i.e. nearly as much as was found by specifically adapted program (PlatPromC) at the same cut off level (81.6%) [1]. In this study performance of PlatPromU was tested on promoters of *H.pylori*, which genome essentially differs from *E.coli* and *C.glutamicum* in terms of GC-content (38.9%, rather than 51-53%). Since most structural modules scored by PlatPromU are AT-rich, this may have negative effect on «*sensitivity*» of unified program.

"*Sensitivity*" of PlatPromU was compared with that of PlatProm and PlatPromH (version adapted to the context of *H.pylori* promoters using 44 experimentally determined transcription start points [3]). The nucleotide sequence of the *H.pylori* genomic DNA was taken from NCBI database (NC_000915). Threshold levels were calculated as suggested [1].

"*Sensitivity*" of PlatPromH appeared to be slightly higher than that of PlatProm (Fig. 1A).

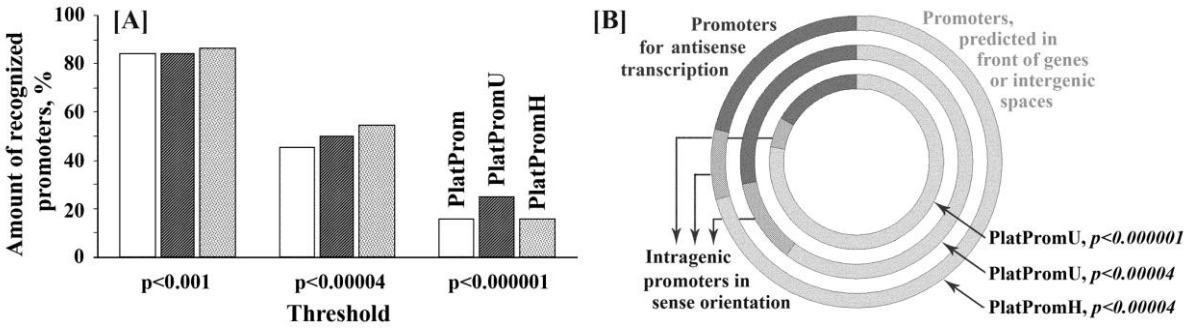


Fig. 1. [A]: An ability of PlatPromH, PlatProm and PlatPromU to recognize *H.pylori* promoters at different levels of reliability. [B]: Classification of predicted promoters in respect to their genomic positioning (%).

About 70% of promoters were predicted in front of coding sequences or within intergenic spaces (Fig. 1B, outer circle). This percentage, as well as portion of intragenic promoters, is practically the same as in the genome of *E.coli* [2]. Efficiency of PlatPromU tends to increase at higher cut off levels (Fig. 1A), while genomic distribution of the predicted transcription signals demonstrates pattern typical for the specific algorithm (Fig. 1B, inner circle). Though the portion of promoters found at this level is rather low (25–33% in *H.pylori* and in *C.glutamicum* [1], respectively), it may be enough to compose learning set required for specific tuning. PlatPromU, thus may be suggested for the preliminary promoter mapping within genomes with uncharacterized regulatory regions, even if their GC-content essentially differs from 50%.

The work was supported by Russian Foundation for Basic Research (grant 10-04-01218).

1. S.S.Kiselev, O.N.Ozoline (2011) Structure-specific modules as indicators of promoter DNA in bacterial genomes, *Mathematical Biology and Bioinformatics*, **6**: 53–65.
2. K.S.Shavkunov, I.S.Masulis, M.N.Tutukina, A.A.Deev, O.N.Ozoline (2009) Gains and unexpected lessons in genome-scale promoter mapping, *Nucleic Acids Res.*, **37**: 4419–4431.
3. C.M.Sharma, S.Hoffmann, F.Darfeuille, J.Reignier, S.Findeis, A.Sittka, S.Chabas, K.Reiche, J.Hackermuller, R.Reinhardt, P.F.Stadler, J.Vogel (2010) The primary transcriptome of the major human pathogen *Helicobacter pylori*, *Nature*, **464**: 250–255.

Human mutagenesis in context

Alexander Panchin[1], Sergey Mitrofanov[2], Sergey Spirin[2,3], Andrey Alexeevski[2,3], Yuri Panchin[1]

[1]*Institute for Information Transmission Problems RAS and Moscow State University, Russian Federation,* alexpanchin@yahoo.com
[2]*Moscow State University, Russian Federation*
[3]*Belozersky Institute of Physical-Chemical Biology, Russian Federation*

A cytosine followed by a guanine (CG) is the most known example of a nucleotide word within the human genome with a dramatically increased probability to undergo mutation. Specific DNA methyltransferase enzymes convert cytosines in CG dinucleotides into methyl-cytosines, which are eventually prone to turn into thymine as a result of deamination. The discovery of other nucleotide contexts that have profound effects on substitution rates can improve our understanding of mutation processes. We compared rates of inherited mutations in 1-4bp nucleotide contexts using reconstructed ancestral states of human single nucleotide polymorphisms (SNPs) from intergenic regions. Chimp and orangutan genomic sequences were used for the purpose of outgroups.

We used a measure called contrast to evaluate if the addition of specific nucleotides to the 5’ or 3’ end of 1–3bp words increases the probability to observe certain mutations in fixed positions. For example, there is a 5.1-fold excess of C to T (C>T) mutations if C is followed by G thus in the position 1 of the word CG. The mutation context description is {CG;C>T;1} and it has a contrast of 5.1 when compared to its {C;C>T;1} subcontext. Contrast values higher than 1 represent excessive mutations while values smaller than 1 represent mutation deficiency. Contrast values for a mutation context {W;mut;pos} and subcontext {W’;mut;pos’} are computed based on the occurrences of words W and W’ and the number of mutations observed in {W;mut;pos} and {W’;mut;pos’}.

Each mutation context can be characterized by two contrast values: mutation bias and minimal contrast. Minimal contrast is the contrast value closest to 1 among all context-to-subcontext comparisons. For example, the {ACG;C>T;2} context has three subcontexts {AC;C>T;2}, {C;C>T;1} and {CG;C>T;1} giving contrast values of 5.08, 5.48 and 1.08 respectively. 1.08 is the minimal contrast for {ACG;C>T;2}. Contrast values obtained using one letter subcontexts such as {C;C>T;1} are called mutation biases. The value 5.48 is the mutation bias for {ACG;C>T;2}.

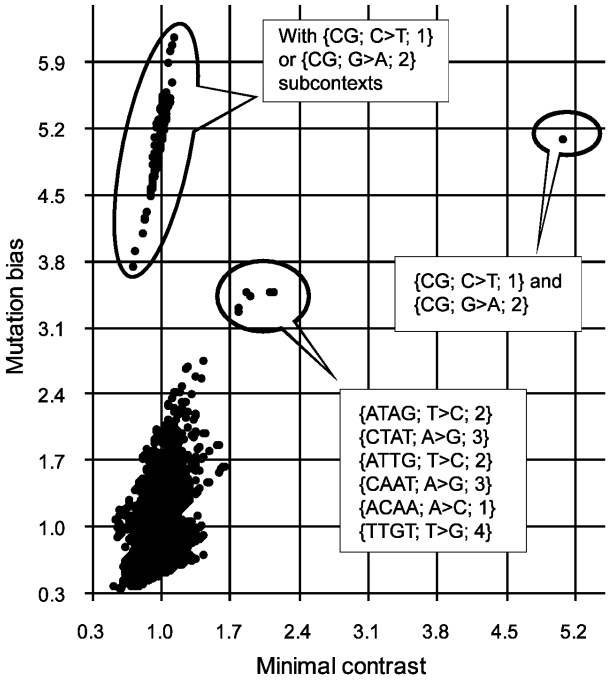


Figure 1. Two-dimensional plot of mutation bias versus minimal contrast of all 1-4bp mutation contexts. Several exclusive clusters are outlined. One contains {CG;C>T;1} and {CG;G>A;2} contexts. It can be discerned by minimal contrast from the another cluster that contains all 3-4bp contexts with {CG;C>T;1} and {CG;G>A;2} subcontexts and only such contexts. The remaining cluster contains six 4bp contexts distinguishable by both minimal contrast and mutation bias.

There are 3.5 and 3.3-fold excesses of T>C mutations in the second position of ATTG and ATAG words respectively and a 3.4-fold excess of A>C mutations in the first position of the ACAA word. Although all observed biases are less pronounced than the 5.1-fold excess of C>T mutations in CG dinucleotides, the three 4bp mutation contexts mentioned above (and complementary contexts) are well distinguished from all other mutation contexts, providing challenges to discover the underlying mechanism that are responsible for the observed excessive mutations.

Nhunt: new program for DNA sequence similarity searching

Yury Pekov[1], Sergei Spirin[2]

[1]*Faculty of Bioengineering and Bioinformatics of Moscow State University, Russian Federation,* yurapekov@gmail.com
[2]*Belozersky Institute of Moscow State University, Russian Federation*

DNA sequence similarity searching in databases is one of the most important problems of bioinformatics. In the case of coding sequences the program TBLASTN is successfully used. But in the case of noncoding sequences, there is no tool satisfying all the needs. A number of programs is used for this purpose, and the most prevailing are FASTA [1], BLASTN [2] and discontiguous MEGABLAST [3]. But each of these programs has some significant disadvantage.

FASTA program shows only one alignment for each database sequence, i.e., the one having the best score among all found alignments. However, in many cases there are a number of regions in a database sequence that are significantly similar to the input sequence.

BLASTN program applies a rapid search algorithm, indexing all words of length N nucleotides (N = 7, 11 or 15) in the entire database. But this approach leads to reduction of sensitivity. If there is no region of N nucleotides in a homologous sequence that is identical to any region in the input sequence, no alignments will be found. Discontiguous MEGABLAST indexes not words, but patterns of length t. A pattern corresponds to a word if it contain at least W (W<t) coincident letters in certain places. This approach is intended to increase the sensitivity, but in reality in most cases the sensitivity even decreases in comparison with BLASTN.

The aim of this work was to create Nhunt computer program for DNA sequence similarity searching that would exceed both FASTA and BLASTN in sensitivity. An original algorithm for diagonals selection was applied, which allows to adjust the ratio "speed / sensitivity". The algorithm for alignment construction is based on FASTA algorithm, but devoid of inherited disadvantages.

Formula for E-value calculation is based on Karlin – Altschul extreme value distribution [4]. Its parameters were fitted using a large random database.

The program was tested on RNA sequences searching in various procaryotic genomes. It was shown that for all examples Nhunt exceeds FASTA program both in sensitivity and in speed. Nhunt also exceeds BLASTN in sensitivity: slightly while searching homologues of E. coli tRNA in a set of archaean genomes, and significantly while searching homologues of E. coli miscRNA in a set of bacterial genomes.

The program is realized on C programming language. Executable files for Linux x86 and amd64 architectures, as well as the source code of the program are accessible in Internet: http://mouse.belozersky.msu.ru/~bennigsen/nhunt.html

The work is partly supported by the grant no. 10-07-00685-a of Russian Foundation of Basic Research.

1. http://faculty.virginia.edu/wrpearson/fasta/
2. ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/user_manual.pdf
3. http://www.ncbi.nlm.nih.gov/blast/discontiguous.shtml
4. Karlin, S. and Altschul, S.F. (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proceedings of the National Academy of Sciences of the USA 87:2264-2268.

Modeling Type I interferon pathway for the study of Multiple Sclerosis

Inna Pertsovskaya[1], Nuria Domedel-Puig[2], Jordi Garcia-Ojalvo[2], Pablo Villoslada[1]

[1]*IDIBAPS- Hospital Clinic of Barcelona, Spain, inna.perts@gmail.com*
[2]*Universitat Politecnica de Catalunya, Spain*

Multiple Sclerosis (MS) is an inflammatory autoimmune neurodegenerative progressive disease. It affects more then 2 million people around the world, mainly young adults (between 20 and 40 years old). The cause of Multiple Sclerosis remains unknown and believed to be both genetic and environmental. The progression of the disease is highly individual and hardly predictable. No treatments exist to cure Multiple Sclerosis. Several drugs are used to modulate the disease and decrease the number and strength of relapses. Interferon beta (IFNbeta) is used in the treatment of multiple sclerosis as a first-choice drug. However, only half of MS patients show a significant positive response to IFN treatment, so the discovering of predictive biomarkers of the response to IFN-beta treatment is extremely important. The mechanism of action of IFNbeta in MS is mainly unknown, but there are evidences that it has antiviral effect, it inhibits the immune cells trafficking across blood-brain barrier, it regulates T-cells activation, etc.

The research of pathway kinetics may help to find new connections between intracellular reactions and response to treatment, and indicate the biomarkers of IFNbeta response in MS patients. Our objective is to develop a kinetic model of the Type I IFN signaling pathway and validate it using experimental data obtained from the experiments with a mouse macrophage cell line. We use the model to predict kinetic alterations in the pathway behavior after IFNbeta therapy of MS between responders and non-responders to the therapy using peripheral blood mononuclear cells from patients and healthy controls.

Our model connects the main components of the JAK-STAT signaling pathway: IFN-beta, isomers of STAT1 and STAT2 proteins and phosphorylated STAT1 and STAT2, inhibitor SOCS1 at both the mRNA and protein level, and the activator IRF1 of STAT1 expression. These

components are connected with positive and negative feedback loops that are responsible for a rich dynamical behavior.

Our experimental data obtained from a mouse macrophage cell line using different molecular biology techniques (RT-PCR, flow cytometry, western blot, luminex, etc.) confirmed the kinetics of the main molecules showed by the model. We confirmed the predicted oscillations of protein concentrations and IFNbeta production during LPS stimulations.

Our first data from PBMCs from blood of MS patients show significant differences in the dynamics of SOCS1 RNA between different individuals. Kinetic changes in the IFN pathway might indicate different response modes to IFNbeta therapy in MS.

Use of hash tables for RNA structure prediction

Dmitri D. Pervouchine[1], Ekaterina E. Khrameeva[1,3], Oleksii V. Nikolaenko[2], Mikhail S. Gelfand[1,3], and Andrei A. Mironov[1]

[1]*Department of Bioengineering and Bioinformatics, Moscow State University, Vorobiovy Gory 1-73, Moscow 119992, GSP-2 Russia, dp@bioinf.fbb.msu.ru*
[2]*Institute of Molecular Biology and Genetics NAS of Ukraine, 150 Acad. Zabolotny, 03143 Kyiv, Ukraine*
[3]*Institute for Information Transmission Problems RAS, Bolshoi Karetny per.19, Moscow, 127994, Russia*

Recent breakthrough in sequencing technology brought in tremendous amounts of sequence data posing new challenges for computational scientists to develop efficient algorithms for RNA structure prediction. Our recent results on the conserved long-range RNA structures associated with splicing have confirmed efficacy and tractability of the novel heuristic approach based on the use of hash tables [1]. There, hash tables were used to find core complementary regions (this is done in linear time by using the reverse complement operation modified to tolerate small number of GT base pairs) in each of the orthologous sequences; the resulting conserved RNA structure is obtained by set-theoretic intersection of hash tables. The output of this procedure is the set of pairs of conserved complementary sequences called 5'- and 3'-boxes.

In the current work we applied a similar method to analyze sequences surrounding splice sites in twelve mammals. Unlike previous work, where only annotated splicing events were analyzed, here we consider all possible combinations of complementary n-mers located in arbitrary combinations of donor (D) and acceptor (A) splice sites (not necessarily ones corresponding to a confirmed intron). There are four possible arrangements of 5'- and 3'-boxes: (DD) both boxes are located next to (possibly different) D-sites, (DA) 5'-box is close to a D-site, while 3'-box is close to an A-site (intron loop-out), (AD) 5'-box is close to an A-site, while 3'-box is close to a D-site (exon loop-out), and (AA) both boxes are located next to A-sites. At that,

the arrangement of boxes in the DD and AA cases can be either in *cis*, where both boxes are located in a neighborhood of the same D- or A-site (*cis*-arrangement corresponds to a local, hairpin-like structure), or in *trans*, where boxes are located in neighborhoods of different splice sites. As a control, we used non-cognate datasets composed of random combinations of sequences neighboring splice sites of different genes. The controls were taken by permutations of hash tables (a) without constraints, (b) with the constraint of having approximately equal GC content, and (c) with the additional constraint of having approximately equal nucleotide conservation rates. Note that, by definition, the total number of cis DD and AA structures in the control procedures is always the same as in the original search and, thus, cis DD and AA structures were excluded from all comparisons and were controlled separately by nucleotide sequence shuffling.

We report that (1) the number of structures predicted in all four arrangement types (DA, AD, trans-DD, and trans-AA) is significantly greater than the corresponding figures in all three controls; (2) the fraction of RNA structures corresponding to annotated splicing events (including events contained in RefSeq and events inferred additionally from EST data) is significantly higher compared to that in controls; (3) alternative splicing events are more frequently associated with conserved RNA structures than are constitutive splicing events; at that, alternative acceptor site usage is the most over-represented category among alternative splicing events;(4) the occurrence of mutually exclusive (DSCAM-like) alternative secondary structures appears to be higher compared to controls; (5) the local structures found in cis-arrangements tend to be associated with alternative donor, but not acceptor site usage; (6) RNA structures are found next to splice sites that tend to be weaker (i.e., further from the consensus in terms of similarity score) than on average. Note that there was no restriction on the distance between 5'- and 3'-boxes (with the exception of cis-DD and cis-AA cases) and, thus, most of the predicted secondary structures were essentially long-range. The estimated false positive rate in the prediction sets did not exceed 40%.

These findings suggest that long-range base-pairing interactions could be essential for the regulation of alternative splicing and demonstrate that long-range RNA structures can be effectively predicted by using hash tables.

The work was funded by RFBR grants 10-04-00783-a and 09-04-92742.

1. Raker VA, Mironov AA, Gelfand MS, Pervouchine DD. (2009) Modulation of alternative splicing by long-range RNA structures in Drosophila *Nucleic Acids Res*. **37(14):**4533-44

Complementing functional annotations and synonyms using cross-species transfer and ortholog mappings

Robert Pesch, Gergely Csaba, Ralf Zimmer

Institut für Informatik, Ludwig-Maximilians-Universität München, Amalienstr. 17, 80333 München, {Robert.Pesch, Gergely.Csaba, Ralf.Zimmer}@bio.ifi.lmu.de

In order to derive and interpret biological networks, a functional annotation of the proteins and genes in the networks is needed. Many approaches have been presented to automate the annotation of proteins in various species [1], but still only a fraction of proteins have a detailed functional description (Fig. 1a). Such functional descriptions can be used to derive synonyms which can be used in text mining and extraction approaches to link literature mentions of proteins and genes to network entities thereby enhancing functional information. Ortholog databases like InParanoid [2] and COGs [3] provide mappings between proteins of different species which can be used to enrich the annotations.

In this study we investigated whether and with which accuracy functional descriptions and synonyms can be transferred between species in order to increase the number of annotated proteins. We propose a method to evaluate the quality of ortholog relations with protein functional descriptions and to transfer ortholog relations transitively along a given phylogenetic tree. For our study, we selected 38 species from InParanoid including the main model organisms and mapped descriptive synonyms like “Interleukin-1 receptor type 1” to the proteins from various synonym databases. For the evaluation of ortholog mappings we extracted function terms from the descriptive synonyms and checked if the function terms overlap from ortholog proteins. Synonyms containing words like homolog, like, or predicted were excluded. The function descriptions of ortholog proteins to Bos taurus have an average consistency of 82% using the direct ortholog transfer and of 83% with the transitive transfer, respectively (Fig. 1b). As the direct mappings are only slightly less accurate then the transitive transfer along the phylogenetic tree (Fig 1b) we can use the more sensitive direct mappings. Protein annotations are transferred using the most reliable relative in the tree and proceeding using more and more distant species. Using this iterative approach, the functional annotation for Bos taurus proteins could be increased from 61% to 82% and 76% using only the species until the first model organism (Fig. 1a) with few expected false transfers (estimated from Fig 1c). The constructed mapping and analysis allows for enhancing the annotations of networks for species.

1. A. Rodrigue et al. (2007) *BMC Bioinformatics*, **8**:S1.
2. G. Ostlund et al. (2010) *Nucleic Acids Research*, **38:**D196-203.
3. R. Tatusov, et al. (2003) *BMC Bioinformatics*, **4:**41.

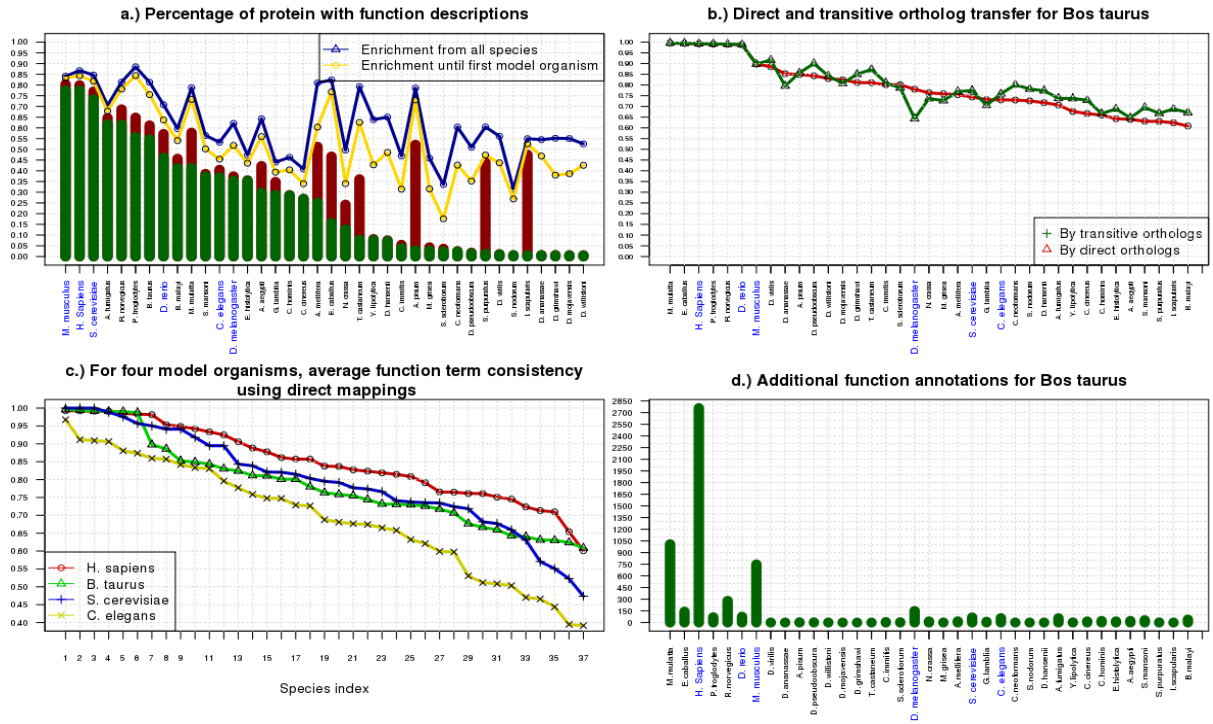


Figure 3 a) *Percentage of proteins with functionally descriptive synonyms. Red indicates the fraction of all proteins with descriptive synonyms whereas green indicates the fraction of all proteins with descriptive synonyms excluding the already transferred synonyms.* ***b)*** *Comparison of the direct and transitive ortholog transfer for Bos taurus: more reliable ortholog mappings show similar accuracy at somewhat reduced sensitivity.* ***c)*** *Average function term consistency for four model organisms using the direct ortholog transfer: the figure gives estimates of accuracies for function description transfer for various distances along the tree of life.* ***d)*** *Additional function annotations inferred for Bos taurus proteins using transfer of descriptions from ortholog proteins using more and more distant relatives according to the phylogenetic tree and the estimated accuracy (Fig 1c).*

Mapping gene expression in two Xenopus species: evolutionary constraints and developmental flexibility

Leonid Peshkin
Harvard University, United States, peshkin@gmail.com

Changes in gene expression are thought to be important for morphological evolution, though little is known about the nature or magnitude of the differences. Here we examine Xenopus laevis and Xenopus tropicalis, two amphibians with very similar development, and ask how their transcriptomes compare. Despite separation for ~30-90 million years there is strong conservation in gene expression in the vast majority of the expressed orthologs. Significant changes occur in the level of gene expression but changes in the timing of expression (heterochrony) were much less common. Differences in level were concentrated in the earliest embryonic stages. Changes in timing were prominently found in pathways that respond to selective features of the environment. We propose that different evolutionary rates across developmental stages may be explained by the stabilization of cell fate determination in the later stages.

De Novo Sequencing of Peptide Antibiotics

Pavel Pevzner[1], Hosein Mohimani[2], Pieter Dorrestein[2], Bill Fenical[3]
[1]*University of California at San Diego, United States, ppevzner@ucsd.edu*
[2]*School of Pharmaceutical Sciences at UCSD*
[3]*Scripps Institute of Oceanography*

Proliferation of drug-resistant diseases raises the challenge of searching for new, more efficient antibiotics. However, sequencing peptide antibiotics, once a heroic effort, remains time-consuming and error-prone. Most antibiotics represent cyclic nonribosomal peptides (NRPs) with nonstandard amino acids that are notoriously difficult to sequence. Moreover, the dominant technique for sequencing antibiotics (NMR) requires large amounts (milligrams) of highly purified materials that, for most compounds, are nearly impossible to obtain. Therefore, there is a need for NRPs sequencing by tandem mass spectrometry from picograms of material. Since nearly all NRPs are produced as related analogs by the same microorganism, we develop a mass spectrometry approach for sequencing all related peptides at once (in difference from the existing approach that analyzes individual peptides). Our results suggest that instead of attempting to isolate and NMR-sequence the most abundant compound, one should acquire spectra of many related compounds and sequence all of them at once using mass spectrometry. We illustrate applications of this approach by sequencing new variants of antibiotics from Bacillus brevis as

well as sequencing a previously unknown family of NRPs (named Reginamides) isolated from a marine bacterial strain that produces natural products with anti-asthma activities.

Inverted Repeats in Surveyed and Sequenced Cattle and Sheep

Anton Pheophilov

Russian State Agrarian University – Moscow Timiryazev Agricultural Academy, foton87@yahoo.com

Nowadays animal genotyping is of great importance because of search for new genetic reserves which may be used in creation of new agricultural breeds. To choose the proper genotyping method a special attention should be paid on reliability, cost and difficulty of preferred marking system. A using of ISSR-PCR method is known to be one of the most convenient from this point of view [5]. According to the method it highlighted regions between certain inverted repeats allowing some conclusions about distribution of the inverted repeats used in those animals. We have analyzed animals of several cattle and sheep breeds by using of di- and trinucleotide microsatellite fragments as PCR primers. Our primers were $(GA)_9C$, $(AG)_9C$, $(GAG)_6C$, and $(CTC)_6C$. Our choice was based on the data acquired about informational capacity of several various microsatellite markers achieved in papers [1, 4].

In cattle primer $(GAG)_6C$ produced 8 amplicons while fragments of 820, 540 and 320 bp length were observed among all animals. Polymorphic loci were presented by 750 bp – found in all Yakutian cattle but only in 9 from 23 of Red Estonian breed; 650 bp – found in half of Black Pied breed and in 22 from 23 of Red Estonian breed. 280 bp length fragment was found only in Yakutian animals. Analyzed samples of Edilbay sheep gave 8 loci (3 were polymorphic) but the differences were not reliable.

Primer $(GA)_9C$ produced only 3 amplicons (590, 510 and 260 bp) with no polymorphism found. It's interesting that in Edilbay sheep the same primer produced a 200 bp amplicon which was able to distinguish intrabreed types [2]. A 290 bp length locus showed differenced in polymorphism compared to Romanov sheep [4]: it's more consolidated in Romanov breed while it's under pressure in Edilbay breed.

The primer $(AG)_9C$ produced 10 amplicons, 4 of them (1000, 650, 500 and 430 bp) were conserved in Edilbay sheep. According to literal data 1000 and 650 bp amplicons are spread also in other sheep breeds [1]. Amplicon with length 1200 bp wasn't described in any investigation analyzed by us and it is likely to indicate breed affiliation. This primer also allowed revealing of intrabreed types in Edilbay sheep: 780 bp fragment was observed only among Birlik type. Amplicons 950 and 280 had a clear correlation with sex of animals.

The last primer, $(CTC)_6C$ produced 17 polymorphic loci but they didn't allow observation of clear interactions with intrabreed type while 2 loci depend on sexual affiliation.

In result, using of ISSR-PCR makes it possible to distinguish species, breed and sometimes intrabreed types by analyzing amplification spectra in sum. A preliminary search in GenBank gives some clues about "sense" of those anonymous DNA. For example, in our recent work we have found relations with MHC class 2 genes, chromatin-associated proteins like GAF and CTCF [3]. Maybe there is connection with intron-exon borders within DNA sequence. However, the role of certain inverted repeats in genome must be investigated further helping us to understand their meaning better.

Author would like to acknowledge dr. V.I. Glazko, dr. T.T. Glazko, dr. Stolpovsky and colleagues in RSAU-MTAA for help and conducting this work.

1. T.N. Dyman et al. (2000) Participation of structure gene markers and anonymous DNA sequences…, *Cytology and genetics*, **6:** 49–59.
2. I.A. Elsukova et al. (2010) Investigation of Suinduk and Birlik intrabreed types of Edilbay sheep…, *Izvestiya of Timiryazev Academy*, **6:** 84–89.
3. A.V. Pheophilov, V.I. Glazko (2010) Structure and functional organization and polymorphism of AG and GA repeats…, *Izvestiya of Timiryazev Academy*, **4:** 104–108.
4. Yu. A. Stolpovsky et al. (2008) Polymorphism of molecular and genetic markers in Romanov sheep breed, *Izvestiya of Timiryazev Academy*, **2:** 48–54.
5. E. Zietkiewicz et al. (1994) Genome fingerprinting by sequence repeat (SSR) – anchored polymerase chain reaction amplification, *Genomics*, **20:** 176–183.

Clusters of Splicing Regulatory Protein Pasilla Are Overrepresented in D. Melanogaster Splice Junctions

Maya Polishchuk[1], James Brown[1], Alexander Favorov[2], Peter Bickel[1]

[1]*University of California, Berkeley, United States, pl.maya@gmail.com*
[2]*Institute for Genetics and Selection of Industrial Microorganisms, Russian Federation*

Pasilla is a Drosophila ortholog of the mammalian NOVA1 and NOVA2 proteins (NOVA), which are well characterized RNA-binding proteins that regulate splicing events in the generation of alternative transcript isoforms [1]. A recent functional study [2] identified hundreds of transcripts that are targets of Pasilla (PS) binding and regulation in Drosophila, and characterized the PS binding motif, YCAY (where Y is any pyramidine). Functional targets of PS regulation were highly enriched for repeats of the YCAY motif (clusters). This study utilized RNAi knockdown of PS. Although this assay revealed hundreds of target, via observations of differential splicing before and after treatment, it is difficult to assess the sensitivity of the

method. A more direct approach would be RIP-seq, Immunoprecipitation of RNA binding proteins followed by high-throughput sequencing and peak calling, but this technique has suffered technical delays in fly. In advance of RIP-seq data, we developed a transcriptome-wide computational approach to predict functional targets of PS based on YCAY motif clusters localization. We used PatternClust [3] to identify regions enriched with YCAY motifs. PaternClust discovers clusters of various lengths and density on the input sequence (whole transcripts), rather than in fixed-size windows.

In FlyBase r5.26 annotations, we discovered a striking pattern: YCAY motif clusters are found in splicing junctions significantly more frequently than in any other transcript location ($p<10-6$). Among clusters intersecting splicing junctions, twice as many intersect 3' exon splice site than 5' exon splice site. We found 47 532 highly significant PS motif clusters ($p<0,007$). Thus, we predict that there are thousands of binding regions for PS as strong and often much stronger than those identified in the functional study [2]. It may be that the largest and most dense motif clusters successfully compete for PS binding even under PS knockdown conditions, and therefore that there are thousands, as opposed to hundreds, of PS regulatory targets in the Drosophila genome, which would be consistent with what has been discovered for NOVA1 in mammals [4].

Thus, our methodology can be used to predict regulatory proteins binding regions, as well as identify potential locations of new exons.

1. J.Ule et al. (2006) An RNA map predicting Nova-dependent splicing regulation, Nature, 444: 580-586
2. A.N.Brooks et al. (2011) Conservation of an RNA regulatory map between Drosophila and mammals, Genome Research, 21:193–202.
3. M.Polishchuk et al. (2008) The binding sites of the proteins regulating transcription in the early development of Drosophila Melanogaster: a comparative analysis of ChIP-chip data and theoretically predicted clusters, Biophysics, 53:352–354.
4. D.D.Licatalosi, et al. (2010) HITS-CLIP yields genome-wide insights into brain alternative RNA processing, Nature, 456:464-469.

Quantitative sequence/activity relationships of auxin response elements (AuxRE) in plant promoters

V.V. Mironova, P.M. Ponomarenko, N.A. Omelyanchuk, M.P. Ponomarenko
Institute of Cytology and Genetics SB RAS, 10 Lavrentyeva, Novosibirsk, pon@bionet.nsc.ru

The hormone auxin is a major regulator of plant growth and development. The auxin action involves the ARF transcription factors that specifically bind to TGTCnC-containing auxin response elements (AuxRE) [Ulmasov et al., 1997, PubMed ID 9188533]. However, using this consensus for ARFs binding sites recognition only gives a huge over-prediction errors.

To find out the additional features of AuxREs we performed the quantitative sequence/activity relationships analysis using data from five types of experiments: (1) the mutation analysis of native AuxREs with either single (14 cases [Ulmasov et al., 1997, , 9188533]) or (2) multiple mutations (21 cases [Ulmasov et al., 1995, 7580254]); (3) the analysis of the AuxRE-like synthetic aptamers (10 cases [Ulmasov et al., 1997, 9401121]); (4) linker-scanning mutagenesis of the auxin responsive 200 bp-region in the *Ps*-IAA4/5 promoter (51 cases [Ballas et al., 1995, 7724586]); (5) the microarray data on auxin-induced expression of the arabidopsis genes with experimentally proven AuxREs (16 cases for 3 microarrays, NCBI).

By generalization of the five dataset we found the obligatory context property accounting for the basic level of auxin induction mediated by the AuxREs [Kolchanov et al., 1998, 9608941]. For this aim we first defined the positional-weight matrix:

$$w(s;i) = \frac{P(s;i)}{P_{MAX}(i) + P_{MAX}(s)} ln\left(\frac{P(s;i)^2}{P_{MAX}(i)P_{MAX}(s)}\right), \quad (1)$$

where: $s \in$ {A,T,G,C}; i – the position on the AuxRE containing [-12;+12] sequence (0 is the C in the tgtCnn consensus); $P(s;i)$ – the probability for the s in the position i; $P_{MAX}(i)$ is the maximum of $P(s;i)$ for position i among all nucleotides; $P_{MAX}(s)$ is the maximum of $P(s;i)$ for the s nucleotide over all positions.

Using eq. 1, the five sets of experimental data were found to correlate with the linear coefficients (1) r=0.709, $\alpha<0.01$; (2) r=0.620, $\alpha<0.01$; (3) r=0.651, $\alpha<0.05$; (4) r=0.407, $\alpha<0.01$; (5) r=0.524, $\alpha<0.05$. In the eq. 1 we empirically introduced the $P_{MAX}(s)$ which accounts for the nucleotide s in the most probable position inhibits occurrence of the s elsewhere. As the properties of the position i and the nucleotides s were equally introduced to the eq. 1, the

assumption of the statistical theory for the DNA/protein binding [Berg, 1987, 3612791] made for the nucleotide position on DNA, could be expanded for the nucleotide type.

Also, we found the facultative properties of AuxREs which may provide for the modulation of auxin response relative to the basic level under different conditions. Using the ACTIVITY program [Ponomarenko et al., 1997, 9109039] we found the conformational, physico-chemical and context DNA properties for each of the five sets that significantly correlate with auxin induced magnitudes in gene expression. Totally, we found 10 facultative properties of AuxREs, per two for each set. Among trinucleotides, the significant linear correlation with the examined IAA-induction magnitudes were found in the cases of TSD ($r=0.823$ at $\alpha<0.01$), HYR ($r=0.612$ at $\alpha<0.01$), VHK ($r=-0.730$ at $\alpha<0.05$), SNW ($r=-0.707$ at $\alpha<0.01$) and VKR ($r=-0.863$ at $\alpha<0.01$), correspondently. In the case of 38 B-helical conformational and physico-chemical properties examined being correlated to IAA-induction we found Tip angle, Helical twist and several others. As an independent control, we have also established the significant linear correlation among the each property and the auxin response in at least one more experiment.

Finally, for the five experiments we described the magnitudes of auxin response by the linear regressions that have the common obligatory AuxRE property (Eq. 1) and the set of specific facultative AuxRE properties. The linear regressions were used for the AuxREs recognition in auxin responsive genes by Poisson's distribution of seldom events combined with Student t-test instead of the commonly accepted recognition threshold.

The work is partially supported by the RAS programs № A.II.5.26, A.II.6.8, B.27.29, SB RAS 107, 119, and RFBR 10-01-00717-a,11-04-01254-a.

Homology Modeling and Comparative Analysis of Serotonin 5-HT3 Receptor Structure in Native and Modified Forms

Anna Popinako

Moscow State University, Biological Faculty, Dept. of Bioengineering, Leninskie gory, 1-12, Moscow, Russia, popinakoav@gmail.com

5-HT3 receptors are members of the Cys-loop superfamily of ligand-gated ion channels which exhibit important physiological functions. Modifications of the 5-HT3 receptors may lead to various pathologies [1,2]. For example, some replacements in extracellular domain lead to increased cardiac activity, increased emotional activity, ultimately lead to stress, anxiety and increased the risk of schizophrenia. [1]

The study of the structure of 5-HT3 receptors is required for the understanding of its role in neurophysiological processes. In this study we present models of 5-HT3 receptors constructed from a homology structure of the nACh, and discuss the amino acid sequence responsible for their different activities.

Molecular models of 5-HT3 receptors were created by applying MODELLER [3]. Some aminoacides residues responsible for the different activities were investigated with the help of the method of molecular dynamics. To study characteristics of the amino acid substitutions of the 5-HT3 receptors we used the GROMACS [4] software with OPLS force field [5].

The created models of the 5-HT3 receptors have demonstrated the abundance of negative charges in the extracellular domain which is seemingly responsible for the direction of the cations migration. It was shown that the steric factor in the region of residue of THR 289 has an influence on the cation transmission. The energy profile analysis has demonstrated the presence of energy minimum in a region that is 2 nm apart from the mouth of the channel. Apparently, it is the region of negative-charged amino acids GLU 272, ASP 293 that take part in a cation hydrate coat reorganization. We observed that ligand binding for the native form of the 5-HT3 receptor is energetically more favorable than those for the modified forms.

The obtained results reveal the relationship between the structure and the different activities of the serotonin 5-HT3 receptors and may be useful in neurophysiological and pharmacological studies.

I would like to acknowledge the advice and guidance of Prof. Konstantin.V. Shaitan, Chair of Bioengineering, Department of Biology, Moscow State University.

1. Andrew J. Thompson, Nora L. Sullivan, Sarah C. R. Lummis (2009) Characterization of 5-HT3 Receptor Mutations Identified in Schizophrenic Patients, *J Mol Neurosci*, **30(3):** 273–281.
2. E.A. Engleman, Z.A. Rodd, R.L. Bell, J.M. Murphy. (2010) The Role of 5-HT3 Receptors in Drug Abuse and as a Target for Pharmacotherapy, *CNS Neurol Disord Drug Targets*, **7(5):** 454–467.
3. N. Eswar, M. A. Marti-Renom, B. Webb, M. S. Madhusudhan, D. Eramian, M. Shen, U. Pieper, A. Sali. (2006) Comparative Protein Structure Modeling With MODELLER. *Current Protocols in Bioinformatics, John Wiley & Sons, Inc., Supplement* **15**, 5.6.1-5.6.30
4. http://www.gromacs.org.
5. George A. Kaminski, Richard A. Friesner E.A. (2001) Evaluation and Reparametrization of the OPLS-AA Force Field for Proteins via Comparison with Accurate Quantum Chemical Calculations on Peptides, *J. Phys. Chem*, **105:** 6474-6487

Weaker selection against internal stop codons in genes with a close paralog in Drosophila melanogaster

Nina Popova[1], Georgii A. Bazykin[1,2]

[1]*Department of Bioengineering and Bioinformatics, Moscow State University, Moscow, Russia*
[2]*Institute for Information Transmission Problems of the Russian Academy of Sciences, Moscow, Russia*
nina.tolmacheva@gmail.com, gbazykin@iitp.ru

After a gene is duplicated, the two copies may diverge in their amino acid sequences in the course of evolution. There has been a considerable debate regarding the different selective pressures driving this process (see Innan and Kondrashov 2009 for a recent review). The majority of the postulated models predict that the negative selection against mutations in each of the two copies increases with sequence divergence, due to a decrease in functional redundancy. We tested this prediction using polymorphism data from 162 D. melanogaster individuals, assessing the occurrence and allelic frequency of internal stop codons in coding regions. Out of 13231 considered genes, stop codons were observed at above-singleton frequencies in 519 genes. The genes with a paralog in the genome were, on average, slightly less likely to carry a stop codon (2.8%) than genes without a paralog (4.1%; chi-square p=0.026). Nevertheless, genes with a closely related paralog carry stops somewhat more frequently than genes with a more distantly related paralog (average protein identity to the nearest paralog: 52.6% for genes carrying a stop codon, 49.7% for genes without a stop codon). However, even for genes with very similar paralogs, stop codons are prevented by substantial negative selection. In summary, our observations support the existing data suggesting that paralogous copies of the gene experience a substantial selection pressure immediately after duplication, and that the selection pressure may be associated with the duplication itself, rather than the subsequent functional divergence.

In silico screening and rational design of multitargeted drugs

Vladimir Poroikov, Alexey Lagunin, Olga Koborova, Olga Filz, Dmitry Filimonov

Institute of Biomedical Chemistry of Rus. Acad. Med. Sci., Moscow, Russian Federation,
vladimir.poroikov@ibmc.msk.ru

Many diseases have a complex etiology, which treatment often requires multiple actions on several pharmacological targets. On the contrary, the majority of current drugs were designed to interact with a single target, which sometimes leads to activation/blockade of other elements in the appropriate signal regulatory pathway. As a consequence of negative feedbacks, expected pharmacotherapeutic effect may be significantly decreased or even completely suppressed. Therefore, the multitargeted drugs, due to their additive, synergistic or antagonistic action, might have some advantages comparing to the monotargeted medicines. The purpose of our study was to develop computer-assisted methods for identification of the most promising targets; finding and rational design of multitargeted agents with the required biological activity profiles.

The following computer-aided tools were used in this work. Net2Drug – software for simulation of behavior of signal regulatory pathways and identification of the most promising targets and their combinations. PASS (Prediction of Activity Spectra for Substances, http://pharmaexpert.ru/passonline) - software, which predicts about 4000 kinds of biological activity on the basis of structural formula with mean accuracy about 95%. PharmaExpert – software for analysis of PASS predicted biological activity spectra and selection of compounds with the required biological activity profiles; GUSAR (General Unrestricted Structure-Activity Relationships) – software for QSAR/QSPR analysis.

As a result, we identified the promising targets for treatment of breast cancers by analysis of signal regulatory pathways. Based on computer prediction of biological activity for 24 mln chemical compounds 64 molecules were selected for experimental testing. 26 samples were purchased and antineoplastic activity was confirmed experimentally in some molecules. Therefore, it was shown that computer-aided methods are rather useful in discovery of the most prospective pharmacological targets and multitargeted ligands.

Acknowledgements. This work was partially supported by European Commission FP6 grant LSHB-CT-2007-037590 (Net2Drug), FP7 grant 200787 (OpenTox), ISTC grants 3197 and 3777.

Structural modeling of BCR-ABL drug resistance mutations

Anna Gorbunova[1], Yuri Porozov[2]

[1]*Saint-Petersburg Pavlov State Medical University, Russia, gorbunova@nm.ru*
[2]*The National Research University of Information Technologies, Mechanics and Optics, Russia, porozov@ifc.cnr.it*

BACKGROUND. The development through a structure based approach of therapeutic agents targeting oncoprotein BCR-ABL has revolutionized the treatment of chronic myelogenous leukemia (CML). Tyrosine kinase inhibitors (TKI) are small-molecule drugs, designed to interfere with BCR-ABL tyrosine kinase by competitive binding at the ATP-binding site in order to inhibit activation of these enzyme. However, mutation-based resistance is an inevitable consequence of kinase inhibitor therapy. Here we report a drug-target complex modeling performed to investigate influence of amino acid substitutions on drug resistance.

METHODS. The Swiss-PDB Viewer [1] and VMD [2] software packages were used to build the models for the complexes of 3 inhibitors (imatinib, dasatinib and nilotinib) with the wild-type and 4 mutants BCR-ABL. Crystal structures of the complex of the kinase domain of wild-type BCR-ABL (inactive form) and inhibitors were used as template structures (imatinib - chain A, PDB code: 2HYY [3]; dasatinib - chain A, PDB code: 2GQG [4]; nilotinib - chain A, PDB code: 3CS9 [5]). Molecular modeling of mutations and binding energy calculation [6] with CHARMm forcefield were performed to study the mechanism of resistance.

RESULTS. Simulations suggested that mutation T315I which demonstrated in experimental data resistance to all analyzed TKIs dramatically increased binding energy in modeled complexes with comparison to wild-type protein (Tab.1). Structural modeling reveals the missing hydrogen bond and steric conflicts are main contributions (Fig.1). The current results demonstrated that for mutation M351T resistance also correlated with calculated binding energy. This correlation was not so clearly seen for Y253H and V299L mutations, and the unfavorable electrostatic interaction between mutated protein and inhibitors may be the main reason for resistance in these mutations.

CONCLUSIONS. Structural modeling of amino acid substitutions in BCR-ABL can predict drug resistance and can help in choosing the best treatment strategy for CML patients. More complex models and additional investigations will be required for some mutations.

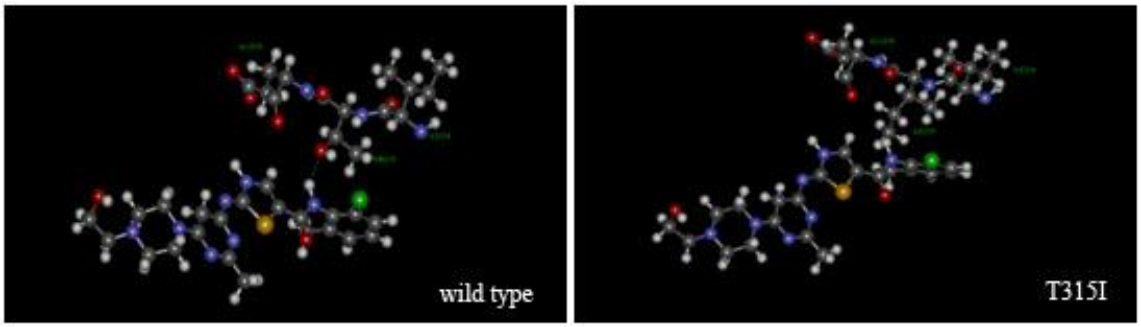


Figure 1. Structural modeling of imatinib binding with wild-type and T315I BCR-ABL.

Table 1. *In silico* calculated binding energy and experimental data on drug resistance for wild-type and mutant forms of BCR-ABL.

Mutation	Imatinib		Dasatinib		Nilotinib	
	Binding energy, (kcal/mol)	Resistance	Binding energy, (kcal/mol)	Resistance	Binding energy, (kcal/mol)	Resistance
wt	-97,39	-	356,14	-	-90,51	-
T315I	2231266,7	+	652095,07	+	106029,12	+
M351T	4107,60	+	356,13	-	-90,16	-
Y253H	-101,24	+	356,50	-	-91,15	+
V299L	4523,49	+	90184,57	+	3302,58	-

REFERENCES.

1. N. Guex and M.C. Peitsch (1997) *Electrophoresis*, **18:**2714-2723.
2. W. Humphrey et al. (1996) *J. Molec. Graphics*, **14:**33-38.
3. S.W. Cowan-Jacob et al. (2007) *Acta Crystallogr D Biol Crystallogr*, **63:**80–93.
4. J.S. Tokarski et al. (2006) *Cancer Res.*, **66:**5790–5797
5. E.Weisberg et al. (2005) *Cancer Cell*, **7:**129–141
6. J. Tirado-Rives and W. L. Jorgensen (2006) *Journal of medicinal chemistry*, **49(20):** 5880-5884.

The structure models of tick-born encephalitis NS2B/NS3 protease for pathogenic and non-pathogenic strains

U.V. POTAPOVA, N.V. KULAKOVA, S. I. FERANCHUK[2],
V.V. POTAPOV, G.N LEONOVA[1], S. I. BELIKOV

Limnological Institute SB RAS, Irkutsk, info@lin.irk.ru, potapova@lin.irk.ru
[1] *Institute of Epidemiology and Microbiology SB RAMS, Vladivostok*
[2] *United Institute of Informatics Problems of the National Academy of Sciences of Belarus, Minsk*

Tick-borne encephalitis virus (TBEV) belongs to flavavirus family. At the moment there is no specific treatment against this decease. Nonstructural protease of flavaviruses (NS3) is responsible for cleavage a protein precursor and is activated by NS2B cofactor. The interest to this protein is explained as it is considered as a potential target for drug design.

The database of TBEV genomes from the patients with subclinical form of decease and isolated from post-mortem tissue [1] was used. On the segment of NS3 protease, there are two mutations between pathogenic and non-pathogenic strains, and near 50% sequence identity to NS3 protease of West Nile viruse.

There are 11 X-ray models of Dengue and West Nile NS2B/NS3 proteases. The models of TBEV proteases were built by homology from the structure with pdb code 3e90 West Nile virus NS2B-NS3 protease as the most complete [2]. Then the molecular dynamics (MD) was performed for both models for 20 ns. The structures remain stable for all the time of simulation.

As the number of residues in both models is the same, the frames of MD trajectories for different strains allow superposition to minimize the average RMSD. Then one can measure the distance between particular correspondent residues between different frames and measure RMSD for a residue between the frames of the same trajectory and between the trajectories as the average over the time of MD run for both trajectories. As it is clear from the plot the two main regions where there is a most significant difference between average positions of the residues in different strains are the segment 35-37 of NS2B and the segment 148-160 of NS3. As it was reported, the residues at NS2B loop at the active site keep their conformation in the long MD simulation and are critical for the catalytic activity of the complex. Our results gives a suggestion that the conformation of the active site is different in the viruses with different pathogenic effect. Typical drug candidates that inhibit protease activity have an active site of the protease and the catalytic triad as the main target. However the substances that inhibit active site of the peptidase may be toxic as they may inhibit other kinds of peptidases, including native

human proteins. The obtained results give a point how to find a site on the protein as a potential target for drug activity, so that the substance will prevent the pathogenic effects of the decease and remains safe as a medicine.

The TBEV protease sequences were taken from database of the GenBank. The sequence alignment of TBEV with West Nile sequence was performed using Genebee [3]. The homology modeling was performed using nest program from jackal package. The explicit water molecular dynamics was performed by Amber11 package. A special program was written for the comparison of MD trajectories from different strains, using W. Kabsch subroutine for structure superposition. The calculations were implemented using framework http://bri-shur.com and packet software Unipro UGENE.

This work was funded in part by Russian Ministry of Science and Education (State Contract No. 14.740.12.0819) and (State Contract No. 389P), Grant ISTC № 4006, Joint project of BRFFR and SB RAS №14.

1. S.Belikov, G.Leonova et al (2010) Coding nucleotide sequences of Tick_Borne Encephalitis virus strains isolated from human blood without clinical symptoms of infection *Genetics*, **3:** 356-363.
2. G.Robin et al. (2009). Structure of West Nile virus NS3 protease: ligand stabilization of the catalytic conformation *J. Mol. Biol*, 385**(5):**1568-1577.
3. V.Nikolaev et al. (1997) Building multiple alignment using iterative analyzing biopolymers structure dynamic improvement of the initial motif alignment *Biochemistry*, **62 (6):** 578-582.

Structural and dynamical properties of human fibrin coiled coil region and its role in the process of fibrin protofibril lateral association

N.A. Pydiura, E.V. Lougovskoy, E.M. Makogonenko, S.V. Komisarenko
O. V. Palladin Institute of Biochemistry, 9 Leontovicha St., 01601, Kyiv, Ukraine, nikkey@bigmir.net

Previously [1] we obtained fibrin-specific monoclonal antibody (mAb) FnI-3C which interacts with fibrin but not with fibrinogen and D-dimer. The epitope for this mAb was localized within region Bbeta Met118-Val134 which is part of the coiled coil region of fibrin(ogen) molecule. MAb FnI-3C as well as its Fab-fragments were shown to specifically inhibit the lateral stage of fibrin polymerization. Furthermore synthetic peptide imitating fibrin(ogen) amino acid sequence Bbeta Leu121-Val138 also inhibits fibrin lateral association while peptide corresponding to Bbeta Gln109-Gln126 doesn't. This gives us strong evidence that this inhibition is implemented by blocking of certain site of protofibril lateral association. We showed that exposition of the epitope for fibrin-specific mAb FnI-3C is result not of distancing or splitting off alphaC domain from coiled coil region but a result of FpA cleavage from fibrin(ogen) molecule. Emergence of the neoantigenic determinant within region Bbeta Met118-Val134 testifies that desAA fibrinogen molecule after FpA cleavage undergoes structural changes which lead not only to protofibril formation but also permits consequent protofibril lateral association. In this study we performed a bioinformatical analysis and structural modeling of the coiled coil region of human fibrin to predict the role of this region in protofibril lateral association and the nature of structural changes that occur during fibrinogen to fibrin transformation.

Bioinformatical analysis of fibrinogen mutations and polymorphism shows the presence of 17 different point mutations and isoforms of fibrinogen, localized in coiled coil region that cause disfibrogenemia with atypical fibrinogen clot formation through inhibited or impaired lateral polymerization. Six of such mutations lay in the region Bbeta Met118-Val138 or close to it: fibrinogen Kyoto IV (Bbeta 111Ser>del), fibrinogen Lyon (Bbeta 118Met>Lys), fibrinogen Epsom (Bbeta 137Asn-141Glu>del), fibrinogen Merivale II (Bbeta 148Lys>Asn). Beside that there are several mutations reported in the corresponding regions of alpha and gamma chains: fibrinogen Plzen (Aalpha 106Asn>Asp), fibrinogen Hannover XVII (gamma 82Ala->Gly).

Consensus antigenicity prediction with four servers of conformational epitope prediction (DiscoTope, BEPro, EPCES, EPSVR) suggests strong antigenic properties for the region Bbeta Ser111-Val138.

To analyze structural changes connected with FpA cleavage the homology model of human fibrinogen was constructed using model PDB code 3GHG as a template. FpAs were modeled and optimized. Procedure was performed in Modeller 9v8 software. Geometry evaluation shows reliability of the final model. Next, dynamical properties of fibrinogen and fibrin X fragments were analyzed with the help of computationally efficient method of correlation molecular dynamics in package CONCOORD 2.1. Obtained trajectories were analyzed by means of analytical tools of Gromacs molecular dynamics software, VMD and PyMOL. Coiled coil region of fibrin fragment X molecule proved to have 40% greater RMSF reflecting an increase in molecular flexibility. Surprisingly coiled coil region containing Bbeta Met118-Val138 fragment showed less critical change in RMSF – 10% increase and 20% decrease of solvent accessible molecular surface. Analysis of the alpha, beta and gamma interchain interactions shows the differences in the spatial organization of coiled coil domain between fibrinogen and fibrin X fragments.

Our results give us the reason to conclude that coiled coil domain plays a crucial role in the process of fibrin lateral association. We approached to localization of the region participating in this process. We suggest that Bbeta Met118-Val138 region is the part of the longitudinal site of protofibril lateral association which is formed as a result of FpA cleavage and consequent increasing flexibility of coiled coil domain. The part of this domain including Bbeta Met118-Val138 contains the unstructured gamma chain region Tyr68-Met78. With the increase of flexibility of the whole fibrin coiled coil domain this part plays the role of a soft hinge remaining less mobile relative to the rest of coiled coil domain.

1. A Lugovskoy E.V., Gritsenko P.G., Kolesnikova I.N., Lugovskaya N.E., Komisarenko S.V. (2009) A neoantigenic determinant in coiled coil region of human fibrin β-chain, *Thromb. Res.*, **123, N5:** 765-770.

Design of specific cytoskeleton related database and data management environment for bioinformatic research in collaboration with virtual Grid-organisation

Nikolay Pydiura, Pavel Karpov, Yaroslav Blume

Institute of Food Biotechnology and Genomics, Natl. Acad. Sci. of Ukraine, Osipovskogo str., 2a, 04123, Kyiv-123, Ukraine, pydiura@gmail.com

Accumulation of vast volumes of biological data led to establishment of a variety of databases. They range from more specialized like sequence database Genebank with 135,440,924 reported sequences as of 15 April 2011 and Protein Data Bank for 3-D structural data with 72550 models as of 19 April 2011, to more integrative databases like UniProt combining protein data from several different databases: Swiss-Prot, TrEMBL and PIR-PSD as well as data derived from literature. But the answer even to a certain simple question requires not only querying of specified databases, and further analysis of the results and integration with derived, object-specific and experimental data. This leads to creation of narrow specific databases, containing information related to specific species, tissues or conditions. Up to date there exist more then thousand of such databases – scientific, medical, agricultural and industrial and its number continues to grow.

Cytoskeletal proteins (tubulins, MAPs, etc) are important targets to a wide range of anticancer, fungicide, herbicide, antiprotozoal, antihelminthic and other agents of commercial importance. In spite of the long history and intense study, understanding of the most of cytoskeletal processes is closely associated with computational (*in silico*) methods. Grid project "Creating a virtual organisation to solve computational problems research cytoskeleton and highly effective screening antimitotic compounds using modern Grid technologies", executed by the Institute of Food Biotechnology and Genomics and sponsored by the National Academy of Sciences of Ukraine (http://grid.nas.gov.ua/) is aimed to involve into collaboration several cytoskeletal groups from Ukraine, Russia, Bulgaria, Czech Republic, Germany, Canada. Virtual organisation CytoLabGrid assumes integration and joint use of computing facilities, databases of molecular structures and experimental results by research groups bound by community of scientific objectives and goals in the cytoskeleton research. The environment should become a rich cytoskeleton related data, information and knowledge repository including genomics, bioinformatics, proteomics and pathways, disease and conditions data capable to provide the possibility of pharmacophore properties prediction, drug design and investigation.

Such virtual Grid-organisation is currently under active development, and its planed to consist of four main parts: 1) query and reporting tool; 2) research data portal; 3) analysis and workflow sharing service; 4) database. Query and reporting tool allows access to the source data from general online databases as well as to the unique derived data computed by system users and stored in the local database which grants its easy retrieval and multiple re-use. Development and implementation of new schemas and pair wise mappings between related schemas grants integrated views adequate to targets. Research data portal provides the interface for linkage of various bioinformatic and experimental data in the research process. Analysis and workflow sharing services provide typical analytical tools, such as for example BLAST, which outputs may be included as parts of the integrated profile. Also it stores the processing information which serves data reliability estimation, automation, bioinformatical knowledge accumulation and scientists collaboration.

System is meant to provide operation with common bioinformatic data types such as DNA and protein sequences, genomics, protein domain organisation , mutations and polymorphism, proteomics, 3-dimentional structure, metabolic and regulaory networks, drug like and pharmacophore small organic molecules, disease and conditions data, literature data. Information resources data annotation and management is based on semantic technologies and imply wide usage of available taxonomies as well as development of new ones.

The designed system will provide an easy access to cytoskeleton elements' data, accumulation of new useful information and its sharing throughout a participating research community.

Acknowledgements: This project is supported by The Ukrainian National Grid - http://grid.nas.gov.ua/ in the frames of State R&D Program "Implementation and Applications of Grid-Technologies"

In SilicoDesigning of an Inhibitor for Initiating the Process of Apoptosis

SUMIT RAJ
SATHYABAMA UNIVERSITY, India, sumitrajde@gmail.com

Beta arrestins are cytoplasmic proteins that bind specifically to active (phosphorylated) G- protein coupled receptors (GPCRs) and arrest or reduce signaling by these receptors. Consequently, it plays a crucial role in cell signaling and various physiological responses. Beta arrestins suppress the GPCR mediated apoptosis. The availability of crystal structure of Beta arrestin-2 with IP6 has offered a great opportunity for homology modeling of Beta arrestins and rationale design of specific Beta arrestins inhibitors. Thus selective Beta arrestins inhibition could be potential therapeutic target for the treatment of cancer and autoimmune diseases. This study is aimed to generate 3-D structures of Beta arrestins and find out the potent and specific Beta arrestins inhibitor by different approaches including homology modeling, molecular docking, virtual screening of ligand databases (drug like) and de novo drug designing. New classes of putative ligands i.e. N- (cyclohexylmethyl) cyclohexanecarbohydrazide were found. After further optimization process these probable drug like molecules can generate a potent inhibitor for the Beta arrestins leading to activation of apoptosis through mitochondrial route.

Protein 3D Structure Prediction by using Heuristics and Structural Restraints

Utkarsh Raj
Amity University, India, rajamity1@gmail.com

Realistic and quick protein structure prediction is crucial for drug design and system biology. There are nearly 30 lacks protein sequences but only 67,000, up to now, protein structures have been experimentally determined and stored in PDB database. There is growing differences between number of experimentally determined 3D structures and Protein primary sequences. Initially ab initio methods were used to predict protein structures but now homology modelling and threading methods are predominant. In Homology modeling, large collection of stored templates is used to predict structure of newly discovered proteins. This technique is very successful in comparison to other Protein structure prediction techniques. Clustering, Rough Sets, Support Vector Machines (SVM) ,Neural Networks ,Case Based Reasoning are some of useful data mining techniques for data mining in bioinformatics. Protein Structure prediction Problem is NP complete and needs heuristic and optimization techniques to find minimum energy structures. Homology Modeling needs fast searching methods for nearest neighbour structural fragments from a database of protein structural fragments. Protein is vital component of all living organism and consist of basically twenty amino acids. Protein is synthesized by DNA in cells of living organism. First DNA creates mRNA and mRNA changes into Protein.

SEQ2GO: Function annotation of hypothetical proteins using sequence based filtering and domain composition of intermediate homologs

Shameer Khader, Sowdhamini Ramanathan
National Centre for Biological Sciences (TIFR), Bangalore, , India, shameer@ncbs.res.in

Bioinformatics-based protocols are often employed for the function association of gene products from genome sequencing project due to the limitation in characterizing the function of gene products using experimental approaches. Homology approaches offers a convenient medium for relating functions [1-3]. Remote homolog searches can be enhanced using sensitive sequence approaches like intermediate sequence search methods. The method was shown to be efficient in connecting in distant relationships through intermediate sequences [4-9]. Association of Gene Ontology (GO) terms to a gene product using integrated approaches are common in the post-genome era as more and more orphan genes are being reported as hits in large scale gene expression and high throughput screening experiments [10-12]. GO annotations remains the most important approach for understanding the function associated with gene products, but due to the limitation in function association using experiment, large amount of currently available annotations are derived from electronic resources and bioinformatics protocols. In this manuscript, we introduce a new GO term association method to annotate hypothetical proteins using sequence based filtering steps and domain composition of intermediate homologs called as "SEQ2GO". SEQ2GO utilizes the concept of intermediate

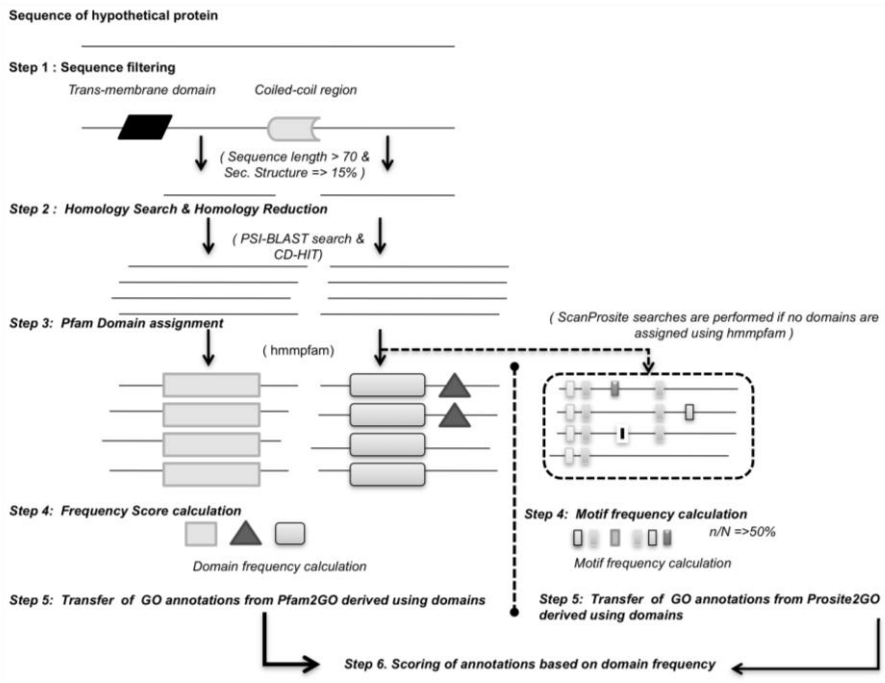


Figure 4: Outline *SEQ2GO method*

sequence searches to assign functional domain to a given unannotate protein with the help of annotations derived from the domain architecture of its homologues (See Figure 1).

SEQ2GO is a multi-step bioinformatics approach that utilise and integrates various bioinformatics tools for the function association of gene products. SEQ2GO is proposed as a method that can be used to assign GO term using an integrated bioinformatics protocol that utilize various sequence features and methods like coiled coil regions, transmembrane regions, secondary structure composition, length criteria, homology search, homology reduction, domain assignment, domain frequency and annotations derived from Pfam2GO and Prosite2GO mappings. The method is broadly divided in to six steps as follows:

Step 1: Sequence filtering - Sequence of hypothetical protein or protein with no GO association is filtered for transmembrane regions using 'tmap' and coiled coil regions using 'pepcoil' from EMBOSS package [13]. Secondary structure calculations are performed using PSIPRED [14] to avoid spurious hits.

Step 2: Homology Search & Homology Reduction - PSI-BLAST search [15] is performed with the individual segments followed by homology reduction performed using CD-HIT within a stringent cut-off.

Step 3: Pfam Domain assignment - Domain assignment of the homologous sequence after CD-HIT search is performed using hmmpfam [16, 17]. Where no domains could be assigned, Prosite pattern search is implemented using ScanProsite [18, 19].

Step 4: Frequency Score Calculation - Individual hmmpfam search results are combined and frequency score calculation is performed for all domains and motifs. Only domains with Frequency score >=50% is used to derive GO terms based on Step 5.

Step 5: Transfer of GO annotations: External2GO (http://www.geneontology.org/external2go/), Pfam2GO and Prosite2GO are used for the GO annotation transfer to hypothetical sequence where there is high frequency score of a domain in homologues.

We applied the SEQ2GO method to an uncharacterized protein in human genome "C2orf16". The gene coding for C2orf16 was reported to be implicated in recent genome-wide association studies on waist Circumference - triglycerides (WC-TG) [20] and serum calcium levels [21]. Following parameters were used: blastpgp e-value=0.001, PSI-BLAST searches were performed using NR as target database, hmmpfam e-value=0.01; CD-HIT threshold=0.4, hmmpfam searches were performed on 50 homologs. SEQ2GO method identified a functional domain "Coagulation Factor V LSPD Repeat (LSPR (PF06049))" in more than 50% of the intermediate homologs with reliable e-values. This domain is used to retrieve the Pfam2GO

annotation term "blood coagulation (GO:0007596). As more and more orphan genes and ORFs are being reported in high-throughput and genome-wide association studies, SEQ2GO can be helpful to delineate and provide biologically relevant clues to the function of such proteins.

1. Copley RR *et al*. *FEBS Lett (*2002), **513**(1):129-134.
2. Lee D *et al*. *Nat Rev Mol Cell Biol (*2007), **8**(12):995-1005.
3. Sowdhamini R *et al*. *Acta Crystallogr D Biol Crystallogr (*1998), **54**(Pt 6 Pt 1):1168-1177.
4. Park J *et al*. *J Mol Biol (*1997), **273**(1):349-354.
5. Salamov AA *et al*. *Protein Eng (*1999), **12**(2):95-100.
6. Reddy CC *et al*. *BMC Bioinformatics (*2008), **9**:281.
7. Reddy CS *et al*. *In Silico Biol (*2006), **6**(5):351-362.
8. Bhadra R *et al*. *Nucleic Acids Res (*2006), **34**(Web Server issue):W143-146.
9. Sandhya S *et al*. *J Biomol Struct Dyn (*2005), **23**(3):283-298.
10. Ashburner M *et al*. *Nat Genet (*2000), **25**(1):25-29.
11. Barrell D *et al*. *Nucleic Acids Res (*2009), **37**(Database issue):D396-403.
12. Rhee SY *et al*. *Nat Rev Genet (*2008), **9**(7):509-515.
13. Rice P *et al*. *Trends Genet (*2000), **16**(6):276-277.
14. Jones DT. *J Mol Biol (*1999), **292**(2):195-202.
15. Altschul SF *et al*. *Nucleic Acids Res (*1997), **25**(17):3389-3402.
16. Eddy SR. *Bioinformatics (*1998), **14**(9):755-763.
17. HMMER: biosequence analysis using profile hidden Markov models [http://hmmer.janelia.org/]
18. de Castro E *et al*. *Nucleic Acids Res (*2006), **34**(Web Server issue):W362-365.
19. Gattiker A *et al*. *Appl Bioinformatics (*2002), **1**(2):107-108.
20. Kraja AT *et al*. *Diabetes (*2011), **60**(4):1329-1339.
21. O'Seaghdha CM *et al*. *Hum Mol Genet (*2010), **19**(21):4296-4303.

From Protein-Protein Interaction Prediction to Elucidation of Missing Metabolic Pathway Enzymes

Vijaykumar Muley, Akash Ranjan

akash@cdfd.org.in
[1]Computational and Functional Genomics Group, Centre for DNA Fingerprinting and Diagnostics, A Sun Centre of Excellence in Medical Bioinformatics, Hyderabad 500001, INDIA

Biological processes are driven by interactions among cellular components. Among such components, proteins are essential and participate in almost every biological process that takes place in cell. Understanding the structure and dynamics of the protein functional and physical interactions is one of the challenging problems in the post-genomic era. Comparative analysis of completely sequenced genomes not only revealed the underlying organizational and evolutionary principles of biological processes but also given a thought that genes should not be studied in isolation [1]. This has led to the development of several high-throughput experimental and computational methods for identification of genome-wide physical and functional protein-protein interactions (PPI) [2-5]. These interactions can be represented in the form of mathematical objects known as graphs (networks) which are composed of nodes and edges. Nodes are the proteins and the edges joining them represent physical or functional interactions [6].

Computational methods have an advantage over experimental methods in terms of cost-effectiveness, high coverage and more importantly, can capture functional interactions. The routinely used computational methods based on genomic context are phylogenetic profile method, a modified gene neighborhood method, referred as minimum distance, gene cluster method and gene order conservation method. Moreover correlated mutations in amino acid sequences of two interacting protein families and correlated expression pattern of genes under various conditions has been effectively used for PPI predictions.

Each method essentially use unique genomic feature for prediction and it has been observed that a number of predicted PPI varied greatly among the genomic context methods. Moreover no method outperforms the others when tested on standard positive datasets [7]. Several studies have shown that the integration of these methods boosts the performance by taking advantage of different genomic features but at the cost of low coverage [4]. Given a high-quality gold standard dataset, machine learning methods (MLMs) can provide a straight forward solution for integration

In this study we use well annotated genome sequence of *Escherichia coli* as a model organisms, to perform in-depth comparative analysis of six PPI prediction methods and their

predictive power. Then we integrate features generated by these six methods and cross-check accuracy of prediction using seven MLMs, which includes Support Vector Machine, Logistic Regression, Naïve Bayes, Bayesian Networks, Random Forest, Decision Tree and Neural Network. Furthermore, we computationally reconstructed genome-wide PPI network of *Escherichia coli* K12 by combining six prediction methods and seven MLMs.

Silent features of our method and analysis are, gene expression and phylogenetic methods outperforms other PPI methods in terms of prediction accuracy and coverage. Positive Prediction Value (PPV) and Specificity of individual MLM is more than 90% and 99% respectively. Though there is not remarkable difference between accuracy of MLM but topological properties of predicted network shows considerable variations. A total of 48381 interactions among 3936 proteins predicted by more than four MLM considered as positives and used to predict missing enzymes in metabolic pathways.

1. Koonin, E.V. and A.R. Mushegian, Complete genome sequences of cellular life forms: glimpses of theoretical evolutionary genomics. Curr Opin Genet Dev, 1996. **6**(6): p. 757-62.
2. Bork, P., et al., Protein interaction networks from yeast to human. Curr Opin Struct Biol, 2004. **14**(3): p. 292-9.
3. Shoemaker, B.A. and A.R. Panchenko, Deciphering protein-protein interactions. Part I. Experimental techniques and databases. PLoS Comput Biol, 2007. **3**(3): p. e42.
4. Shoemaker, B.A. and A.R. Panchenko, Deciphering protein-protein interactions. Part II. Computational methods to predict protein and domain interaction partners. PLoS Comput Biol, 2007. **3**(4): p. e43.
5. Hu, P., et al., Global functional atlas of Escherichia coli encompassing previously uncharacterized proteins. PLoS Biol, 2009. **7**(4): p. e96.
6. Barabasi, A.L. and Z.N. Oltvai, Network biology: understanding the cell's functional organization. Nat Rev Genet, 2004. **5**(2): p. 101-13.
7. Sun, J., et al., InPrePPI: an integrated evaluation method based on genomic context for predicting protein-protein interactions in prokaryotic genomes. BMC Bioinformatics, 2007. **8**: p. 414.

Comparative genomics based reconstruction of transcription regulation network in Staphylococcaceae

Dmitry Ravcheev, Dmitry Rodionov

Sanford-Burnham Medical Research Institute, La Jolla, California, USA, Institute for Information Transition Problems, Moscow, Russia, dravcheev@sanfordburnham.org, rodionov@sanfordburnham.org

Transcriptional regulatory networks are fine-tuned systems, which help microorganisms respond to the changes in the environment and cell physiological state.

Using various comparative genomics techniques [1] implemented in RegPredict web-server [2] we reconstructed regulatory network in the human pathogen *Staphylococcus aureus* and six related species from the *Staphylococcaceae* family. The resulting reference set of 46 transcription factor regulons contains more than 1,900 binding sites and 2,800 target genes involved in the central metabolism of carbohydrates, amino acids and fatty acids, respiration, stress response, metal homeostasis, drug and metal resistance and virulence. The inferred regulatory network in *S. aureus* includes ~320 regulatory interactions between 46 transcription factors and ~550 candidate target genes comprising 20% of its genome. In the reconstructed *S. aureus* regulatory network, we predicted ~170 novel interactions and 24 novel regulons for the control of the central metabolic pathways. The reconstructed regulons are largely variable in the *Staphylococcaceae*: only 20% of *S. aureus* regulatory interactions are conserved across all studied genomes. Available expression data allowed the assessment of the reconstructed regulatory network in *S. aureus*.

All predicted regulons are captured in the RegPrecise database [3].

This is joint work with Pavel S. Novichkov, Andrey L. Osterman and Aaron A. Best.

1. D.A.Rodionov (2007) Comparative genomic reconstruction of transcriptional regulatory networks in bacteria, *Chem. Rev*. .**107**:3467–3497.
2. P.S.Novichkov et al., (2010) RegPredict: an integrated system for regulon inference in prokaryotes by comparative genomics approach. *Nucleic Acids Res.*, **38**:W299–W307.
3. P.S.Novichkov et al., (2010) RegPrecise: a database of curated genomic inferences of transcriptional regulatory interactions in prokaryotes. *Nucleic Acids Res.*, **38:**D111–D118.

T-Rex: Redox-sensitive regulation of hydrogen production in Thermotogales

Dmitry A. Ravcheev[1,2], Dmitry A. Rodionov[1,2]

[1]*Sanford-Burnham Medical Research Institute, La Jolla, California, USA,*
[2]*Institute for Information Transition Problems, Moscow, Russia*
dravcheev@burnham.org; rodionov@burnham.org

Nicotinamide adenine dinucleotides play important role in different biological processes, including redox cellular balance. Previously the Rex protein was described in a number of bacteria, such as *Streptomyces coelicolor* [1], *Bacillus subtilis* [2], *Staphylococcus aureus* [3], and *Thermus aquaticus* [4]. Rex protein is a unique transcription factor that senses the NADH:NAD^+ ratio. Under low NADH:NAD^+ ratio, Rex protein binds to the target sites and represses transcription of genes involved in anaerobic respiration. Increase of the NADH concentration to more than 2% of common NAD(H) pool results the dissociation of the Rex from DNA [1-4].

The deep-branched order of Thermotogales includes thermophilic bacteria isolated in extreme environments from geothermally heated marine sediments to hot springs that can produce hydrogen by fermenting a wide range of carbohydrates. Fermentative hydrogen production is catalyzed by multiple iron-containing hydrogenases that oxidize NADH generated during glucose catabolism. In the previous study of the model species of *Thermotoga maritima* we revealed 18-bp palindromic DNA motif (GK-box) associated with genes involved in glycerate metabolism (*sat-hpr*, *TM1586-gckA*), glycerol utilization (*glpFKPZ*, *gldA*) and glycolysis (*dnaX-TM0687-gap-pgk-tpiA*) [5]. Recently, 11 genomes of different Thermotogae species were sequenced and annotated, and we have analyzed them with the aim of reconstruction of novel transcription regulons using the comparative genomics approach and the RegPredict web-server. Using the previously defined GK-box, we inferred a novel regulon that was significantly expanded by novel target operons such as hydrogenase operons (*TM1420-21-22-23-hydCAB-rex1*, *hycD*, *hycABC*, *hycEC2*), alcohol dehydrogenase (*aldH*), NADH:polysulfide oxidoreductase (*naoX*), and hypothetical CRIPR system (*TM1814-07*). Most of the identified novel operons have conserved candidate regulatory sites in the *Thermotoga* genus, and at least seven of these operons (including *gap-pgk-tpiA*, *hyd*, *hyc*, *aldH*, and *naoX* genes) have the predicted GK-box regulatory sites, which are well conserved in the *Thermosipho*, *Fervodobacterium*, *Petrotoga*, and *Kosmotoga* genomes.

Using the phyletic pattern of GK-box-sites and the genome context analysis of predicted

target genes we assigned two putative paralogs from the Rex family of transcription factors, TM1427 and TM0169, as candidates to function as regulators of the novel GK-box regulon. In support of this assignment, the identified GK-box motif in Thermotogae resembles the previously known Rex-binding motifs in Firmicutes. Phylogenetic analysis of Rex proteins revealed that TM0169 and its orthologs in all Thermotogae genomes form a group of highly conserved proteins (termed T-Rex), that are closely related to the Rex proteins in Firmicutes, whereas the less conserved TM1427 group (termed Rex1) always co-occurs and clusters on the chromosome with the *hydCAB* genes. We proposed that T-Rex is a novel global regulator of hydrogen production and central metabolism in Thermotogae, whereas Rex1 is a local regulator of a single Fe-hydrogenase operon.

To assess the validity of *in silico* genomic–based regulon reconstruction, both T-Rex and Rex1 proteins from *T. maritima* were cloned, purified and tested for binding to their predicted target DNA operator sites using electromobility shift and fluorescence polarization assays. As result, all 14 predicted binding sites of T-Rex in *T. maritima* have been confirmed in the *in vitro* assays. Screen for candidate T-Rex effector molecule has identified that T-Rex-DNA binding is inhibited by NADH and, in a less extent, by NADP, whereas NAD+ and NADP+ do not have this negative effect. In contrast, binding assays did not reveal a binding of Rex1 protein to the T-Rex operators. Phylogenetic footprinting analysis revealed additional conserved region in DNA upstream of the *TM1420-21-22-23-hydCAB-rex1* operon. Candidate Rex1-binding site with consensus sequence, which is weakly similar to the established T-Rex motif, was identified in this region and is a currently experimentally validated.

This is joint work with Andrei L. Osterman and Xiaoqing Li (Sanford-Burnham Institute, La Jolla California, USA), and Vasiliy Portnoy (University of San Diego, California, USA)

1. D.Brekasis, M.S.Paget (2003) A novel sensor of NADH/NAD$^+$ redox poise in *Streptomyces coelicolor* A3(2), *EMBO J.*, **22:**4856-4865.
2. E.Wang et al. (2008) Structure and functional properties of the *Bacillus subtilis* transcriptional repressor Rex, *Mol. Microbiol.*, **69:**466-478.
3. M.Pagels et al. (2010) Redox sensing by a Rex-family repressor is involved in the regulation of anaerobic gene expression in *Staphylococcus aureus*, *Mol. Microbiol.*, **76:**1142-1161.
4. K.J.McLaughlin et al. (2010) Structural basis for NADH/NAD+ redox sensing by a Rex family repressor, *Mol. Cell*, **38:**563-575 .
5. C.Yang et al. (2008) Glycerate 2-kinase of *Thermotoga maritima* and genomic reconstruction of related metabolic pathways, *J. Bacteriol.*, **190:**1773-1782.

Prediction and Validation of Plant DYRK1A Homologues Spatial Structure

Alex Rayevsky, Pavel Karpov, Maxim Korablyov, Stanyslav Isayenkov, Yaroslav Blume
Institute of Food Biotechnology and Genomics, Natl. Academy of Sciences of Ukraine, Ukraine,
rayevsky85@gmail.com

Dual-specificity tyrosine phosphorylation-regulated kinases or DYRKs (Dual specificity YAK1-related kinases), play a key role in the signaling pathways regulating nuclear functions during cell proliferation and differentiation. It is well known that phosphorylation of tau microtubule-associated proteins by Dyrk1A, related to brain development **[**PMID: 17906291**]** and Down syndrome [PMID: 18405873]. The DYRK family kinases are autophosphorylated on tyrosine, serine and threonine residues, but their catalitical activities are strongly associated with serine and threonine phosphorylation [PMID: 8631952].

Previously, we have identified the row of plant homologues of animal microtubule- and cell cycle related serine-threonine PKs [1]. The goal of this study was reconstruction of spatial structure of plant DYRK-homologues and their catalytic functions. Accordingly to PDB-BLAST search we specified template X-ray PDB-structures (2VX3, 3ANQ, 1Z57) for template-based protein structure modeling (Modeller (9v8)). Energy minimizations of 3D-models were produced *in vacuo* and then in the water (model Tip3p) by Charmm27 force field (www.charmm.org). The accuracy of models (human Dyrk1A and two plant homologues) was verified by the MOLprobity server analysis (http://molprobity.biochem.duke.edu/). In order to understand the role of active site conservative residues, the docking of ATP-competitive inhibitors: EHB ((1z)-1-(3-ethyl-5-hydroxy-1,3-benzothiazol-2(3H)-ylidene)propan-2-one) and D15 (N-(5-{[(2S)-4-amino-2-(3-chlorophenyl)butanoyl]amino}-1H-indazol-3-yl)benzamide) was performed. We have reconstructed the complexes of plant DYRK homologues with and ADP as well. All docking simulations were performed in Autodock 4.0 (http://autodock.scripps.edu/) and Dock 6 (http://dock.compbio.ucsf.edu/) programs. GROMACS (v.4.5.3) molecular dynamics simulations were carried out by charmm27 force field (www.gromacs.org/). The ligand models were built in Marvin Sketch (www.chemaxon.com) editor, and topology files were generated by SwissParam server (www.swissparam.ch).

The Y321 phosphorylation (pY321) is necessary for DYRK1A (PDB: 2VX3) activation were shown [PMID: 21126318]. Taking in to account this fact, we have identified similar conserved pY-residues in plant homologs. pY-residues were reconstructed by Marvin Sketch

editor. The ESP charges were calculated by RESP ESP charge Derive (http://q4md-forcefieldtools.org). The every reconstructed system was inserted in a water box with 5 Å water layer, and subjected to steepest descent energy minimization (20,000 steps). Protein backbone was frozen and solvent molecules with counter ions were mobile during a 100 ps position restrained MD run. All simulations were run under periodic boundary conditions, with NPT ensemble by using Berendsen thermostat (at constant T=310 K° and P=1 bar), and 2 fs time step simulations. Electrostatic interactions were calculated by particle-mesh Ewald (PME) algorithm, at 0.12 nm grid spacing and fourth order interpolation. Van der Waals forces were treated using a cutoff of 10 Å and the coordinates were stored every 100 ps. Subsequently, free dynamics simulations were performed at 310 K° for 5 ns.

The average RMSDs of ligands are around 0,04 nm for EHB and 0,1 nm for D15. The SR columbic interaction energy between site and EHB, D15 are -84,3 kJ/mol and -221,4 kJ/mol respectively. The slight differences for average number of H-bonds between active site and the substrates were observed. The protein models from *Physcomitrella* and *Rattus* exhibit 1.8 bonds for EHB and 3.5 for D15. The *Arabidopsis* protein model demonstrates the worse average values (1,08 bonds for EHB and 2.5 for D15).

Finally, we have confirmed structural homology of human Dyrk1A and its plant homologues and laid the basement for further research of their catalytic activity.

1. Karpov P.A., et al. (2009) Bioinformatic Search of Plant Microtubule- and Cell Cycle Related Serine-Threonine Protein Kinases, *Proceedings of the International Moscow Conference on Computational Molecular Biology (MCCMB'09)*. July 20-23, Moscow, Rossia: 145-147.

Computing the *p*-values of selections in huge sets

Jeremie Bourdon[1], Mireille Regnier[2]

1LINA, CNRS UMR6241, Nantes, France, Jeremie.Bourdon@univ-nantes.fr
2AMIB-Inria team, LIX-Ecole Polytechnique, 91 128 Palaiseau, France, mireille.regnier@inria.fr

Introduction. Results from experiments at a genome-wide level [ST03] often are highly noisy. Statistical methods provide insightful results. In the case of multiple testing, several methods have been designed to correct the computed *p*-values, allowing a selection of sub-parts of elements having particular characteristics. The main difficulty there is to provide a suitable threshold, one important criterion being that the number of false positive must be kept under control. Two classical extremal methods exist [Sha95]: Bonferroni correction and Benjamini-Hochberg. We provide a complete framework, both theoretical formula and efficient algorithm that allow dealing with genome-wide data, for computing efficiently and accurately the number of false positive of a given selection. These methods are then applied to ChIP-chip data and sequence analysis.

Methods. ***Probabilistic model***. The data is composed by M independent experiments, for which a binary decision (positive/negative or presence/absence) is made. The overall number of false positive is thus a sum of M independent Bernoulli random variables with non-equal parameters p_i. The Gaussian approximation may hold under smooth conditions, but leads to wrong results in the extremal cases: exact values or large deviation results are required.

Large deviation result. We denote by $p_{k,M}$ the probability that at most k elements amongst M have the correct decision. Let now $0<a<1$ and $\sigma_j = \Sigma_i (p_i)^j$. We establish the following large deviation result enabling an accurate estimation of $p_{Ma,M}$:

$$\lim_{M\to+\infty} \frac{-\log p_{Ma,M}}{M} \sim at_a - (a-\sigma_1)[1 - \frac{\sigma_2(\sigma_1 - a)}{\sigma_1^2}], \text{ where } \quad t_a \sim \log[1 + \frac{a-\sigma_1}{\sigma_1-\sigma_2} + \frac{(a-\sigma_1)^2(\sigma_2-\sigma_3)}{(\sigma_1-\sigma_2)^2}]$$

Dynamic algorithm. A dynamic algorithm, based on some basic recurrences, can also easily deduced for computing $p_{k,M}$. Its theoretical complexity, $O(M^2b)$, also depends on the required number of precision digits b:

$$b = \max\{-\sum_{i=1}^{M} \log_{10}(1-p_i), -\sum_{i=1}^{M} \log_{10} p_i\}$$

Discussion. In a ChIP-chip experiment [CMS+06] on estrogen receptor ESR, around 10000 promotor sequences have been annoted as "positive". This entails a large number of false positives and false negatives. In order to select the sequences where ESR1 binds, up to a fixed false positive frequency, possible strategies (summarized in the following table) are (1) keep the

sequences with the k best p-values; or (2) use p-value adjustments and a given threshold (say 1%). The binding site of ESR1 is now included in some public databases (motif MA00112 in JASPAR). We used it to compute the maximal p-value of ESR1 motif in all the sequences (thus obtaining M p-values p_i). Finally, we determine the smallest k such that $p_{k,M}>0.99$, leading here to $k=3791$.

Strategy 1	Best 10	Best 100	Best 3791
False positive rate	2.10-806 %	9.10-653 %	1%
Strategy 2	Bonferroni	Benjamini-Hochberg	MPV
Sequences selected False positive rate	0	692 3.10-183 %	3791 1%

Conclusion. Our method applies whenever one has to decide between two states according to some probability and extends when more than two states are possible (co-occurrences).

References

[CMS+06] J. S. Carroll, et al.. (2006) *Genome-wide analysis of estrogen receptor binding sites*. Nat. Genet., **38:**1289–1297.
[Sha95] J P Shaffer. (1995) *Multiple hypothesis testing*. Annual Review of Psychology, **46(1):**561–584.
[ST03] J D Storey and R Tibshirani. (2003) *Statistical significance for genomewide studies*. PNAS, **100(16):**9440–9445 August 2003.

Novel Non-Coding Organism-specific Regulatory RNAs

Isidore Rigoutsos

Computational Medicine Center, Thomas Jefferson University, United States, Isidore.Rigoutsos@jefferson.edu

Through earlier computational work we discovered the "pyknon" class of DNA sequence motifs that exhibited a number of intriguing properties. These motifs led us to postulate the existence of specific, previously unseen categories of short RNAs and of an associated framework of putative interactions in which these RNAs participated. Experimental work and additional computational analyses by us and others have begun to provide support for the validity of the pyknon framework and suggest the possibility that a potentially significant portion of cellular process regulation may be mediated by genomic sequences that need not be conserved across organisms.

Integrative reconstruction of carbohydrate utilization metabolic pathways and regulatory networks in Thermotogales

Dmitry RODIONOV[1,2], Vasiliy PORTNOY[3], Xiaoqing LI[1], Irina RODIONOVA[1], Dmitry RAVCHEEV[1,2], Andrei OSTERMAN[1]

[1]*Sanford-Burnham Medical Research Institute, La Jolla, California, USA;*
[2]*Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia;*
[3]*University of California San Diego, California, USA*
rodionov@burnham.org

Bacteria of the deep-branched genus *Thermotoga* can produce hydrogen by fermenting a wide range of carbohydrates. A remarkable diversity of the Thermotogales *sugar diet* is matched by a large fraction of genes committed to carbohydrate degradation and utilization in their genomes. As a result the evolutionary plasticity of the sugar catabolic machinery in general, and due to a unique taxonomic position and lifestyle of Thermotogales, exact functions of many respective genes (and pathways) remained unclear even in the best-studied model species, *T. maritima*. To address this problem we applied an *integrative subsystems-based approach to the genomic reconstruction of metabolic and regulatory networks*. This approach established and validated in our previous studies (e.g. in the reconstruction of sugar catabolic machinery of the Shewanella genus) included three major levels of integration: (i) comparative analysis of 11 complete genomes from the Thermotogales order annotated by RAST server; (ii) parallel genomic reconstruction of biochemical transformations, uptake mechanisms and transcriptional regulation, and (iii) combining bioinformatic predictions with experimental testing in *T. maritima* model. Application of various bioinformatics tools implemented in the SEED (theseed.uchicago.edu) and RegPredict (regpredict.lbl.gov) web-sites allows us to substantially improve the accuracy of annotations as well as predict novel, previously uncharacterized genes and pathways.

The *genomic encyclopedia of sugar utilization* in Thermotogales includes more than 300 functional roles spanning at least 20 distinct pathways with mosaic distribution across 11 analyzed species. The detailed results of this analysis are captured in the subsystem "*Sugar utilization in Thermotogales*" available online from the SEED web-site. The current version of the subsystem comprises >130 cytoplasmic and extracytoplasmic sugar catabolic enzymes (including ~40 glycoside hydrolases), ~90 components of carbohydrate uptake systems (mostly ABC transporters), and 18 committed transcription factors. The developed metabolic model incorporates all published experimental data and inferences about enzyme activities, substrate specificities of transporters, and differential gene expression patterns on various carbohydrates (generated mostly for *T. maritima*). Our analysis revealed substantial differences in sugar

catabolic pathways between Thermotogales and other previously studied bacteria. Most common are *nonorthologous gene replacements*, when a functional role is encoded by a gene, which is not orthologous (and, often, nonhomologous) to any of the previously described genes of the same function. The repertoire of transporters and regulators involved in sugar catabolism in *Thermotoga* demonstrates the most prominent differences in comparison with other taxa. We validated two novel pathways for utilization of inositol and galacturonate and characterized 15 carbohydrate kinases in *T. maritima* (details are on the Rodionova et al poster). Finally, the predicted catabolic capabilities of *T. maritima* were assessed by monitoring growth rates, substrate consumption and gene expression on a panel of various individual and mixed mono- and disaccharides.

A transcriptional regulatory network inferred from comparative genomic analysis of Thermotogales includes 32 transcription factors and their DNA binding sites unevenly distributed across 11 studied genomes. A current collection of regulons captured in the RegPrecise database (regprecise.lbl.gov) is centered on *T. maritima* and includes 18 transcription factors that were predicted to control expression of ~185 genes involved in sugar catabolic machinery of this model organism. Remarkably, a large fraction of these genes and operons are controlled by multiple transcription factors pointing to a complexity of regulatory responses to changing environmental conditions (Fig. 1). For example, we established partial overlaps between the xylose, glucuronate, and galacturonate regulons (XylR, KdgR, and UxaR, respectively); the glucose, trehalose and inositol regulons (GluR, TreR, and IolR); and the cellobiose, mannose, and glucooligosaccharide regulons (CelR, ManR, and GloR). The experimental assessment of the reconstructed regulatory network included *in vitro* analysis of selected individual regulons and *in vivo* gene expression profiling of *T. maritima* on various carbohydrate substrates. We used the first approach based on gel-shift mobility assays to validate all predicted DNA targets and identify small molecule effectors for six regulators from the ROK family (BglR, IolR, XylR, ChiR, TreR, and ManR). We are currently expanding this effort to characterize additional transcriptional regulators, Rex, UxaR, KdgR, CelR, and RhaR. Global gene expression profiles were obtained and analyzed for the growth on 12 different carbon sources using high-density oligonucleotide tiling arrays (Nimblegen). Gene induction patterns measured for tested mono- and disaccharides (trehalose, rhamnose, xylose, etc.) showed a strong correlation and provided additional information to refine respective regulons (TreR, RhaR, XylR, etc.) reconstructed by the genomic analysis.

Comparative genomics approaches for reconstruction of transcriptional regulatory networks in Bacteria

Dmitry A. RODIONOV[1,2], Pavel S. NOVICHKOV[3]
[1]*Sanford-Burnham Medical Research Institute, La Jolla, California, USA,*
[2]*Institute for Information Transition Problems, Moscow, Russia;*
[3]*Lawrence Berkeley National Laboratory, Berkeley, California, USA*
rodionov@burnham.org, psnovichkov@lbl.gov

Genome-scale annotation of regulatory features of genes and reconstruction of transcriptional regulatory networks in a variety of diverse microbes is one of the critical tasks of modern Genomics and Systems Biology. A growing number of complete prokaryotic genomes allow us to extensively use comparative genomic approaches to infer *cis*-acting regulatory elements (e.g. transcription factor binding sites and riboswitches) in regulatory networks of numerous groups of bacteria. Two major components of this analysis are *propagation* of previously known regulons from model organisms to others and *ab initio prediction* of novel regulons, as implemented in the integrative web-server tool RegPredict (http://regpredict.lbl.gov).

We developed and utilized the integrative comparative genomics approach to infer transcriptional regulatory networks in ~100 microbial genomes from ten distinct taxonomic groups of bacteria (each including from 6 to 16 genomes): *Shewanella*, *Desulfovibrionales*, *Enterobacteriales*, *Cyanobacteria*, *Thermotogales*, *Bacillales*, *Streptococcus*, *Staphylococcus*, *Ralstonia*, and *Corynebacteria*. A limited input of established regulon members was provided by publications on particular transcription factors in individual species (*E. coli*, *B. subtilis*, *S. aureus*). The reconstructed regulatory networks for the key pathways involved in central metabolism, production of energy and biomass, metal homeostasis, stress response and virulence includes over 450 regulators, >22000 their DNA-binding sites, and 25 families of RNA regulatory elements (e.g. riboswitches). The obtained reference set of microbial regulons is captured in the RegPrecise database within the taxonomic group-specific collections (http://regprecise.lbl.gov). Many novel regulons first predicted and reconstructed by the comparative genomics techniques were validated by targeted in vivo and in vitro experiments. The obtained by genomic analysis network of regulatory interactions provides a framework for the interpretation of gene expression data in model species.

Characterization of novel components of sugar catabolic pathways in *Thermotoga maritima* identified by integrative genomic approach

Irina A. RODIONOVA[1], Dmitry A. RODIONOV[1,2]

[1]*Sanford-Burnham Medical Research Institute, La Jolla, California, USA;*
[2]*Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia,*
irinar@burnham.org; rodionov@burnham.org

The marine hyperthermophilic bacterium *Thermotoga maritima* has extensive and highly diversified carbohydrate utilization machinery including polysaccharide breakdown, uptake mechanisms, biochemical transformations in the cytoplasm and transcriptional regulation. Accurate functional assignment of this machinery is challenging due to substantial variations of the respective pathways between species including frequent nonorthologous gene displacements and functionally divergent paralogs. Therefore, their homology-based annotations in various genomic databases are often incomplete and imprecise. To address this challenge we combine a comparative genomics subsystems-based approach (implemented in the SEED genomic platform, http://www.theseed.org) with the experimental reconstitution of pathways using *in vitro* enzymatic assays, high-performance liquid chromatography (HPLC), and gas chromatography–mass spectrometry (GC-MS). First, two novel pathways for utilization of inositol and galacturonate have been predicted in the *Thermotogales* genomes and experimentally validated in *T. maritima*. Second, we inferred and experimentally assessed substrate specificities within the *T. maritima sugar kinome* represented by at least 15 sugar kinases involved in a variety of carbohydrate utilization pathways.

A novel variant of the galacturonate/pectin utilization pathway is encoded by the TM0430-43 gene cluster in *T. maritima*. It encodes extracellular pectin hydrolytic enzymes, digalacturonate ABC transporter, and cytoplasmic enzymes including exopolygalacturonase PelB, gluconate-6P dehydrogenase Gnd, and four novel enzymes (named UxaE, UxaI, UxaD, and GntK) presumably forming a novel pathway of transformation of galacturonate to gluconate-6P. We purified these novel enzymes, as well as UxaC and UxuB enzymes encoded in the glucuronate utilization gene cluster, and validated their predicted activities using *in vitro* biochemical assays. UxaC is a bifunctional enzyme catalyzing isomerization of glucuronate to fructuronate and galacturonate to tagaturonate shared by the two respective pathways. UxuB is a NAD-dependent mannonate dehydrogenase. UxaE catalyzes the epimerization of tagaturonate to fructuronate, a novel reaction that was not previously described in any organism. UxaD is a bifunctional NADPH-dependent 5-keto-gluconate/fructuronate reductase. GntK is a gluconate-

specific kinase. The galacturonate and glucuronate catabolic gene clusters were found to be co-regulated by a novel GntR-like regulator (UxaR) that was also validated by *in vitro* DNA-binding assays.

A novel variant of the inositol catabolic pathway in *T. maritima* is encoded by the TM0411-TM0422 gene locus. The first gene encodes a novel ROK-family transcription factor (IolR) that was validated to bind to its operator site in the promoter region of the inositol operon. The InoEFGL ABC transporter was proved to bind inositol or inositol-1-phosphate. We purified and characterized biochemically the conventional inositol dehydrogenase IolG, and three novel enzymes (named IolM, IolN, and IolK) that presumably form the novel inositol pathway. The first two steps are catalyzed by the known inositol dehydrogenase IolG and the novel inosose dehydrogenase IolM (K_m for NAD = 0.14 mM and K_m for inosose = 0.7 mM). 5-ketogluconate identified as a product of the IolN-catalyzed reaction is further utilized using a novel reductase (UxaD) from the reconstructed galacturonate catabolic pathway in *T. maritima*. The resulting gluconate is phosphorylated by a novel kinase (IolK), and further metabolized via the pentose-phosphate pathway.

Sugar phosphorylation is an essential step in the carbohydrate catabolic pathways in bacteria. Whereas in many bacteria this step is often performed by uptake-associated phosphotransferases (PTS), in *T. maritima* it appears to be fully delegated to the members of its extended and diversified *sugar kinome*. Using genome context analysis and metabolic reconstruction we assigned specificities to nearly all sugar kinases in *T. maritima*. We used a "matrix" approach to experimentally test these assignments and to explore the relationship between the inferred physiological roles and *in vitro* substrate preferences of respective enzymes. Purified recombinant proteins were tested for their kinase activity versus a panel of >40 different mono- and disaccharides. Remarkably, nearly all of the 15 experimentally characterized enzymes (mainly of from the FGGY and PfkB families) displayed a strong preference towards a single physiological substrate. The ROK-family kinase Glk was found to possess broad substrate specificity to hexoses: glucose, mannose, N-acetylglucosamine, and glucosamine. Overall, the reported results illustrate the efficiency of the subsystems-based approach and, specifically, the choice of sugar kinases as signature enzymes for recognition and reconstruction of sugar utilization pathways.

This is joint work with Andrei L. Osterman (Sanford-Burnham Medical Research Institute, USA).

Comparison of quality and performance of parallel algorithms for Multiple Sequence Alignment

Kirill Romanenkov, Alexey Salnikov

faculty Computational Mathimatics and Cybernatics of Lomonosov Moscow State University, Russia,
kromanenkov2@yandex.ru

The parallel algorithms of sequence processing are increasing in popularity from year to year. This is caused by about 2 times increase in size of the biological databases per year. The speed of perfection for specific software isn't so impressive. Considering the fact that many manufactures now accentuate on the multicore architectures and growth the number of cluster's nodes the problem of fast and accurate parallel algorithm is rather essential. This article illustrates features of parallel algorithms for multiple sequence alignment.

Several parallel algorithms (modification of sequential MUSCLE with PARUS system [1], Dialign-P [2], ClustalW-MPI [3]) have been tested for accuracy and scalability. MUSCLE and ClustalW-MPI are the representatives of progressive alignment family which characterizes by smaller time of work and worse quality of alignment because it depends on the initial clustering. Dialign-P falls into the category of iterative algorithms. This class is defined by more accurate alignment because it allows realign of already proceed sequences if the score function doesn't take an optimal value. Of course, this feature extremely increase the work time even parallel iterative algorithms. The sequences of all Long Terminal Repeats class 5 from human genome (1500bp x 1200 seq.) and thirteen protein families from Pfam database (526bp x 1100seq.) were used for performance comparison on multiprocessor configurations using from 1 to 32 processors. The results of testing showed that Dialign-P demonstrates the worst execution time in the group also its' rather exacting for the amount of memory per each HPC cluster's node. For this reason it was impossible to run it on some data which provide large size of algorithm's input. MUSCLE showed good scalability but on small amount of sequences displays inadequate reduction of execution time. ClustalW-MPI demonstrated better scalability but provided worse execution time.

Accuracy of the algorithms was compared with align a special set of sequences called "multiple sequence alignment benchmarks": BAliBASE, OxBench, SABmark. Quality of built alignments was counted with Q-(amount of correct aligned pairs/amount of pairs in standard alignment) and TC-score(amount of correct aligned columns/amount of columns in standard alignment) metrics. Results have shown that parallel algorithms, in common, don't lose in quality

in comparison with their sequential analogs. In some cases difference in the achieved quality was connected with the fact that parallel algorithms did not find the common parts in sequences. The closest result to its sequential analogue showed ClustalW-MPI. The best score was achieved by Dialign-P, which seems predictable because this algorithm belongs to the iterative methods. The algorithms scalability was investigated by comparison of the execution time of the parallel and the corresponding sequential algorithm for alignment of diverse nucleotide or amino acid sequences. It was demonstrated that good performance of parallel algorithm usually corresponds with a low quality of output, while algorithms that provided more accurate output were not scalable and required a large amount of memory per each HPC cluster's node.

1. Alexey N. Salnikov The modification of MUSCLE multiple sequence alignment algorithm for multiprocessors Proceedings of the 3-rd Moscow conference on computational molecular biology, Moscow, Russia, July 27-31 2007, pp. 270-271.
2. Martin Schmollinger, Kay Nieselt, Michael Kaufmann and Burkhard Morgenstern DIALIGN P: Fast pair-wise and multiple sequence alignment using parallel processors // BMC Bioinformatics, 2004, 5:128, ISSN: 1471-2105.
3. Kuo-Bin Li ClustalW-MPI: ClustalW analysis using distributed and parallel computing //Bioinformatics Vol. 19, No. 12, 2003, pp. 1585-1586. ISSN: 1460-2059 (Electronic),ISSN: 1367-4803 (Print).

An algorithm for exact probability of pattern occurrences calculation

Evgenia Furletova[1], Mireille Regnier[2], Mikhail Roytberg[1], Viktor Yacovlev[1]

[1]*Institute of Mathematical Problems of Biology, 4, Institutskaja str., 142290, Pushchino, Moscow Region, Russia*
mroytberg@impb.psn.ru
[2]*INRIA, 78153 Le Chesnay, France*

An important aspect of studying functional fragments of biological sequences is to determine the statistical significance of their occurrences. One of the statistical significance measures is P-value i.e. probability to find at least p occurrences of a pattern (a set of words) H in a random sequence of length n. We have created an algorithm SufPref to calculate the P-value. We assume that words from the pattern have the same length (pattern length).. The algorithm calculates P-value for three types of probability models: Bernoulli, Markov models of order K, Hidden Markov models. The program that implements the algorithm is available at http://server2.lpm.org.ru/bio

Unlike the majority of existing programs that calculate exact P-value only for Markov models of order 1 or two, our program supports Markov models of any order less then length of words in the pattern. As to our knowledge at the moment there are no program computing exact P-value for Hidden Markov models. In the Bernoulli case both of the time and space complexities of SufPref are independent of the alphabet size and the time complexity in average is independent of the pattern length. We have compared our algorithm with algorithm Spatt based on minimal finite automaton [1]. In most of test examples the running time of SufPref is better than one of SPatt, especially in the cases where patterns are randomly generated. The space complexities of both algorithms are compatible. A detail analysis of complexity of the algorithm for Bernoully and Markov cases and its comparison with other algorithms are given in the paper [2].

1. G. Nuel. Effective p-value computations using Finite Markov Chain Imbedding (FMCI): application to local score and to pattern statistics. Algorithms for Molecular Biology 2006, 1:5 doi:10.1186/1748-7188-1-5
2. M. Regnier, Z. Kirakosyan, E. Furletova, M. Roytberg. An word counting graph. London Algorithmics 2008: Theory and Practice, 2009.

The influence of intron length on the intron phase distribution

Tatiana Astakhova[1], Ivan Tcitovich[2], Mikhail Roytberg[1]

[1]*Institute of Mathematical Problems of Biology, 4, Institutskaja str., 142290, Pushchino, Moscow Region, Russia, mroytberg@impb.psn.ru*
[2]*Institute of Information Transmission Problems, RAS, Moscow, Russia*

We have studied the dependence of intron phase distribution on intron length (intron phase is the number of nucleotides in an incomplete codon preceding the intron) in various species.

Species	Pos	40-200		5000+		T	p
		N1	F1	N2	F2		
Apis mellifera	1	1060	0,34	249	0,42	2,53	5,8E-03
	2	1363	0,32	147	0,41	2,36	9,1E-03
	3	1241	0,32	94	0,40	1,66	4,8E-02
	4	1025	0,30	68	0,44	2,46	7,0E-03
Drosophila. Melano-gaster	1	3222	0,33	761	0,45	6,17	3,9E-10
	2	2726	0,29	216	0,40	3,18	7,3E-04
	3	2143	0,29	145	0,39	2,48	6,7E-03
	4	1561	0,29	142	0,49	5,15	1,5E-07
Anopheles gambiae	1	785	0,31	201	0,39	2,09	1,9E-02
	2	530	0,25	117	0,36	2,30	1,1E-02
	3	388	0,27	67	0,30	0,53	3,0E-01
	4	268	0,28	60	0,37	1,31	9,6E-02
Nasonia vitripennis	1	1514	0,36	268	0,44	2,38	8,8E-03
	2	1630	0,31	131	0,36	1,11	1,3E-01
	3	1393	0,29	91	0,35	1,21	1,1E-01
	4	1253	0,31	63	0,35	0,62	2,7E-01
Tribolium castaneum	1	1978	0,35	314	0,46	3,67	1,2E-04
	2	1536	0,31	147	0,36	1,37	8,5E-02
	3	1225	0,31	104	0,40	2,03	2,1E-02
	4	946	0,30	59	0,37	1,19	1,2E-01

Table 1. Legend: Pos - position of an intron in the gene, N - number of introns of the specified length in the 1 st phase; F1 - the frequency of introns in the 1 st phase among all introns of the specified length; T - value of Student's t test, p – confidential probability.

It was shown that for insects the frequency of introns in phase 1 increases for "abnormally long" introns (length ≥ 5000, most introns are about 80 bp), see Tab.1. In [1] it was shown that for different taxa the frequency of introns in phase 1 depends of the position of the intron in the gene. Therefore we separately considered introns in different positions. The effect holds for all 4 positions of introns holds, but most strong is for the first intron.

This effect is most significant for D. melanogaster, however, it is observed in all considered insects. The same effect can be observed also for “medium length” introns (200-1000 and 1000-5000 bp), see Fig.1.

For other taxa situation differs from the insects' one. The effect was not observed (*Ciona_intestinalis)* or was very weak (*Cearnohabditis elegans* and *Xenopus Silurana tropicalis).* In genome *Hydra magnipapillata* one can observe the opposite effect (the frequency of introns in phase 1 decreases with icreasing of introns length), in genomes, *M_musculus, Macaca mulatta, Homo sapiens* and plants (*Populus trichocarpa, Arabidopsis thaliana , Oryza_sativa, Vitis vinifera)* this opposite effect can be noted only for the first introns. Detailed information can be found at http://lpm.org.ru/~mroytberg/Introns_phases.

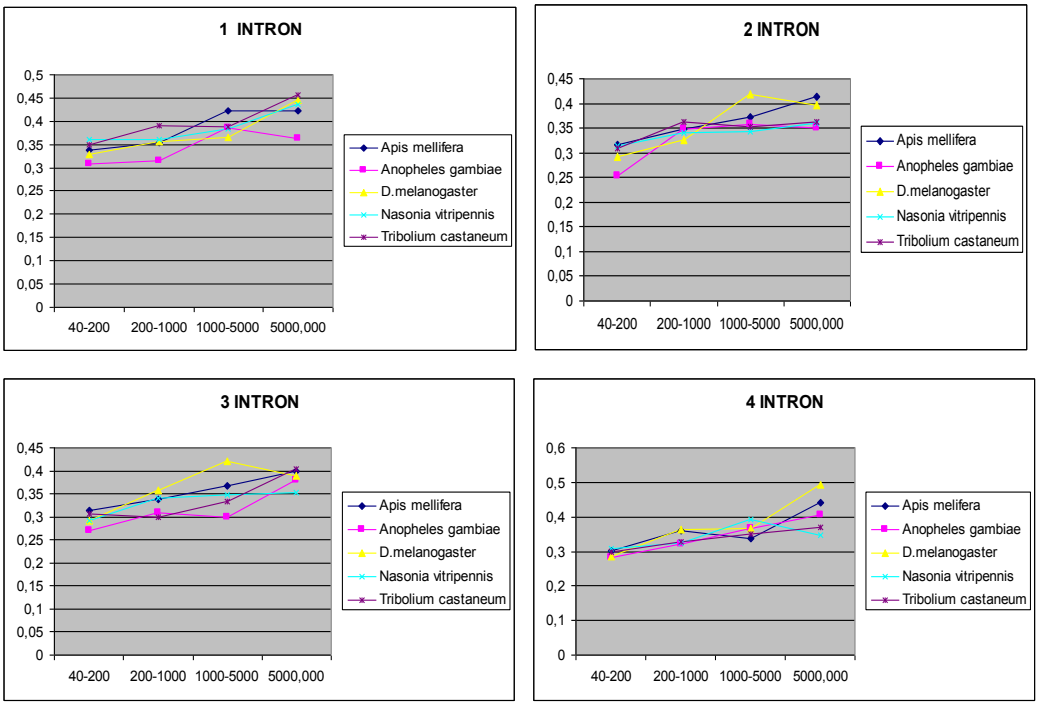


Fig.1. Frequency of phase 1 introns for various species, intron lengths and intron positions

1. A. Ruvinsky, W. Ward A Gradient in the Distribution of Introns in Eukaryotic Genes. (2006) J Mol Evol 63:136–141.

Comparative analysis of genomes of 12 species of Drosophila

Tatiana Astakhova[1], Dmitry Malko[2], Vsevolod Makeev[2], Mikhail Roytberg[1]

[1]*Institute of Mathematical Problems of Biology, Russian Academy of Sciences, Puschino, Russia,* *mroytberg@impb.psn.ru*
[2]*Vavilov Institute of General Genetics RAN 119991, Moscow,Gubkina,3, Russia*

We have performed comparative analysis of neighborhoods of exon borders in complete genomes of genus *Drosophila.* For each exon border we have picked up two-side neighborhood, length of each flank equals 38 bp. Thus we have considered 8 types of fragments (coding and non-coding flanks of conservative dinucleotide of acceptors sites, the same for donor sites, start- and stop-codons). The dataset was prepared using programs [1] and Pro-Frame [2].

Following characteristics were calculated for each type of fragments *t*, position *i* and genome *g*: frequency $PN[t, i, x, g]$ of each nucleotide *x;* frequency $PD[t, i, xy, g];$ of each dinucleotides *xy;* Likelihood ratio $LR[t, i, xy, g] = PD[t, i, xy, g]/(PN[t, i, x, g] \cdot PN[t, i, y, g])$ of each dinucleotides *xy;* for each type of fragments we have calculated also mean values of the characteristics (notation: $MPN[t, x, g], MPD[t, xy, g], MLR[t, xy, g],$. All data are available at http://lpm.org.ru/~mroytberg/DrosophilaSITE.ZIP.

The most significant results are as follows. (1) Frequencies of A and T nucleotides in exonic flanks in *D.willstoni* genome is greater than in other genomes, see Tab. 1, 2. It is known [3] that the preference of codons in *D. willistoni* differs significantly from other types of Drosophila, it is consistent with our data. (2) Likelihood ratio of dinucleotide GC in the genomes of *D. mojavensis, D. virilis, D. grimshawi* is greater than in other genomes; (3) Likelihood ratio for conventional conservative dinucleotide AG in the intron neighborhood acceptor sites is lower than 1 for all genomes, this is in correspondence with the known data for other species. (4) Likelihood ratio for dinucleotide TA in all exonic neighborhoods is lower than 1 for all genomes.

1. W.J..Kent. BLAT -.(2002) Genome Research, vol. 12: 656-664.
2. A.A.Mironov, P.S.Novichkov, M.S.Gelfand. (2001) Bioinformatics, 17:13-15.
3 Saverio Vicario, Etsuko N Moriyama and Jeffrey R. Powell. Codon usage in twelve species of Drosophila, (2007) *BMC Evolutionary Biology*, **7:**226

a b

	A	C	G	T	SPECIES	A	C	G	T
DMEL	26.26	22.86	11.18	39.71	DMEL	24.64	27.19	26.14	22.03
DSIM	26.10	23.16	11.10	39.64	DSIM	24.39	27.51	26.22	21.89
DSEC	26.09	23.24	11.15	39.52	DSEC	24.18	27.70	26.30	21.83
DYAK	26.00	23.42	11.05	39.53	DYAK	24.25	27.65	26.50	21.60
DERE	25.68	24.30	11.82	38.20	DERE	24.24	27.97	26.48	21.32
DANA	26.52	21.55	9.76	42.17	DANA	24.77	27.40	25.77	22.07
DPSE	21.53	26.04	12.38	40.05	DPSE	24.18	27.18	26.64	21.99
DPER	22.11	25.13	12.26	40.51	DPER	24.21	27.07	26.38	22.33
DWIL	23.44	19.00	9.57	48.00	DWIL	26.49	23.37	24.08	26.05
DMOJ	24.47	21.05	10.45	44.03	DMOJ	25.27	25.05	25.44	24.23
DVIR	25.25	19.50	11.75	43.49	DVIR	25.35	25.40	25.73	23.53
DGRI	25.68	18.38	11.31	44.62	DGRI	25.13	24.50	25.93	24.44

Tab.1. Average nucleotide composition of intronic (a) and exonic (b) neighborhoods of acceptor sites for different species of *Drosophila.*

a b

SPECIES	A	C	G	T	SPECIES	A	C	G	T
DMEL	26.52	26.59	25.49	21.40	DMEL	31.30	19.68	18.54	30.48
DSIM	26.04	26.88	25.64	21.45	DSIM	31.13	20.05	18.82	30.00
DSEC	26.16	26.95	25.67	21.22	DSEC	30.91	20.04	18.93	30.11
DYAK	26.05	27.18	25.57	21.19	DYAK	30.65	20.40	18.39	30.56
DERE	25.82	27.37	25.88	20.93	DERE	29.91	21.11	19.27	29.72
DANA	26.05	26.83	25.18	21.94	DANA	30.16	19.72	17.90	32.22
DPSE	26.42	26.56	25.52	21.50	DPSE	27.59	24.90	20.76	26.76
DPER	26.68	26.07	25.28	21.97	DPER	28.58	24.14	20.23	27.05
DWIL	29.42	21.65	23.15	25.78	DWIL	35.17	18.02	15.91	30.89
DMOJ	27.51	25.14	24.51	22.85	DMOJ	33.10	19.30	17.35	30.24
DVIR	27.24	25.23	24.87	22.65	DVIR	33.06	20.86	16.67	29.41
DGRI	27.67	25.49	24.03	22.81	DGRI	34.28	19.29	15.03	31.40

Tab.2. Average nucleotide composition of exonic (a) and intronic (b) neighborhoods of donor sites for different species of *Drosophila.*

Statistics of RNA structures

Evgeny Baulin[1], Dmitriy Ivankov[2], Mikhail Roytberg[3]

[1]*Higher School of Economics, Russian Federation*
[2]*Technical University of Munich, Germany*
[3]*Institute of Mathematical Problems in Biology RAS, Russian Federation, mroytberg@lpm.org.ru*

We present first results of investigation of experimentally obtained RNA structures. The aim of the study is to learn more about special type of bonds between RNA nucleotides ("links", see below).

The initial data were taken from Protein Data Bank [1] and Nucleic Acid Database [2]; only documents containing only one model we have considered. That gave us 10300 documents describing structures with RNA, the documents divided into 4 categories: RNA; RNA-Protein complex; RNA-DNA complex; RNA-DNA-Protein complex. The option "find_pair" from the X3DNA [3] were used to create the files containing the information on bonds between nucleotides and on the number of helices, that constitute these bondings. At this stage 52 original files were temporary put aside, because they included modified elements that cannot be handled with X3DNA program. The results of the "find_pair" operation were converted to tables, which became the subject of the further processing. At the current stage of investigation we are interested in the RNA structures themselves; therefore we have temporary put aside 2621 files containing several RNA bound with each other.

We use the following terminology: D-helix – Helix defined by X3DNA; Isolate – Bonding defined by X3DNA as isolated; Standard helix (Helix) – Non-extendable sequence of X3DNA-bondings of form: (x, y), (x+1, y-1),....,(x+s, y-s), here (i, j) denote the bond between i-th and j-th nucleotides of the chain. Link – Bond that is not a part of the standard helix of length 2 or more.

We say that bonds (a, b) and (c, d) are *correlated* if the fragments (a, b) and (c, d) do not intersect each other or one of them is a part of another. Otherwise we say that the bonds are *conflicting*. Since a nucleotide cannot form more than one bond a link conflicts with a bond belonging to a standard helix iff it conflicts with all bonds of the standard helix. Therefore we say "a link conflicts with a helix", "two helices are conflicting", and so on. The links have been divided in three classes: Internal (non-conflict with spirals or other links), Connected (with conflicts only with other links), Free (with conflicts with helices). At this stage 900 files containing conflicts between the helices were temporary put aside.

The analysis of the data have resulted in tables of several types (table of helices, table of unpaired fragments ("loops"), table of links, etc.). Our study for example shows that ~ 36% of considered structures contain links. About 30% of links have conflicts with helices, about 55% of links has no conflicts at all and only 15% are "connected" links, i.e. links having conflicts with other links but not with helices. Interestingly, positions of nucleotides forming a link are close to ends of helices. As to our knowledge the role of links in RNA structure was not studied so far. At the next stages of investigation we plan to learn more about types of loops related to links, relation between links and RNA-protein interaction and so on.

We thank S.A.Spirin for useful discussions.

1. http://www.pdb.org/pdb/home/home.do
2. http://ndbserver.rutgers.edu
3. http://rutchem.rutgers.edu/~xiangjun/3DNA/

Using of PREFAB for analysis of amino-acid sequence alignment algorithms.

Irina Poverennaya[1], Mikhail Lobanov[2], Victor Yacovlev[3], Mikhail Roytberg[3]

[1]*Moscow State University, Russian Federation*
[2]*Institute of Protein Research RAS,, Russian Federation*
[3]*Institute of Mathematical Problems in Biology RAS, Russian Federation, mroytberg@lpm.org.ru*

Reference alignments are essential for correct analysis of amino-acid sequence alignment algorithms, because their comparison with algorithmic alignments allows one to assess the quality of these methods. The protein reference alignment benchmark PREFAB [1] was used in many studies (e.g., see [2, 3]); it contains 1682 alignments which were obtained using 3D alignment of protein structures. Unfortunately, selection principles for aligned sequence pairs aren't described. At the same time, sequence names include only PDB ID and chain and it remains unclear what fragments of this chain were used.

The aim of our study is to find out correspondence between PREFAB sequence pairs and SCOP structural domain classification [4]. Briefly, we accept amino-acid sequence, if (1) it is a fragment of one chain in a PDB entry [5], and (2) this fragment match with at least one of domains which were describeded in SCOP. An alignment is accepted only if domains corresponded compared sequences belong to the same «family» of SCOP classification.

The analysis of PREFAB have shown the following results: PREFAB has been proved to include sequences that form domain with fragments from other chains of the same protein or contain more than one domain. Besides some chain fragments were found deleted in some one-domain PREFAB sequences. These fragments are likely to correspond with unstructural regions of protein. Consequently we select 1294 sequences (or 1115 PREFAB alignments) that satisfy mentioned conditions. 834 alignments of them satisfy the condition related to SCOP families. A quantity of insertions and sequence identity were computed for each selected alignment. In addition, new structural alignment was built for each PREFAB alignment using the program [6] and distances between the corresponding residues were calculated.

The refined dataset will be used to assess quality to alignments obtained with a method [2] (server address: [7]).

1. PREFAB benchmark: http://www.drive5.com/muscle/prefab.htm
2. Yacovlev V.V., Roytberg M.A. Increase of global amino-acid sequence alignment accuracy by alignment-candidate set construction // Biophysics. - 2010. - T. 55, N 6. - C. 965-975
3. Edgar, Robert C. (2004), MUSCLE: multiple sequence alignment with high accuracy and high throughput, Nucleic Acids Research 32(5), 1792-97.
4. Murzin A. G., Brenner S. E., Hubbard T., Chothia C. (1995). SCOP: a structural classi_cation of proteins database for the investigation of sequences and structures. J. Mol. Biol. 247, 536-540.
5. Protein Data bank: http://www.pdb.org/pdb/home/home.do
6. Structural alignment: http://phys.protres.ru/~mlobanov/casp/prog.html#MaxSub
7. Server: http://server2.lpm.org.ru/bio/online/pareto/

Analysis of distance matrices and construction of phylogenic trees

Pavel Perevedentsev[1], Mikhail Roytberg[2], Sergei Spirin[3]

[1]*Higher School of Economics, Russian Federation*
[2]*Institute of Mathematical Problems of Biology RAS, Russian Federation, mroytberg@lpm.org.ru*
[3]*Moscow State University, Russian Federation*

Quality of phylogenetic tree critically depends on the quality of distant matrix used by an algorithm constructing the tree [1]. Is it possible to reveal features of distance matrices allowing to choose the best one from the list of candidates or to improve the given matrix in way allowing improving the quality of resulting phylogenetic tree? As to our knowledge the question was not considered so far.

The approach we used is based on the well-known "4-leaves" feature of tree distances. Let A, B, C, D be arbitrary four leaves of a weighted unrooted tree and distance d(X, Y) between two leaves is the sum of weights of edges connecting X and Y. Consider three possible partitions of the 4-tuple {A, B, C, D} into two pairs and three corresponding sums d(A,B)+ d(C, D); d(A, C)+ d(B, D); d(A, D)+ d(B, C). Let SP0, SP1, SP2 be the sums ordered by increasing. The "4-leaves" equality claims that SP1=SP2 > SP0.

This allows one to formulate criteria of "tree-likelihood" of the distances between 4 leaves A, B, C, D given a distant matrix. Let SP0, SP1 and SP2 be the above sums. We have considered following quality factors: Q1 = SP2-SP1; Q2 = (SP2-SP1)/SP2; Q3 = (SP2-SP1)/)(SP1+SP2)/2); R1 = ((SP1+SP2)/2) -SP0; R2 = ((SP1+SP2)/2 -SP0)/SP0; R3= ((SP1+SP2)/2 -SP0)/((SP1+SP2+SP0)/3). The Q-factors reflect the needed coincidence between SP1 and SP2; the R-factors reflect the distance between nearest common predecessors of "paired" leaves.

To check predictive abilities of the factors we have performed computer experiments based on the set of bacterial genes, the set contains 103 families of aligned orthologous genes from 30 bacteria; the "true" phylogenetic tree for the bacteria was determined from the integral information about all genes. For each of 103 gene families we have constructed 3 distance matrices based on given multiple alignments: B-matrix (based on BLOSUM62 weight matrix), M-matrix (based on identical weight matrix) and J-matrix (based on Jones-Taylor-Thornton distance). Then starting from each of the 309 matrices we have constructed phylogenetic trees using neighbour-joining (NJ) algorithm. For each of the matrix and each 4-tuple of bacteria we have computed the following values: (1) values SP0, SP1, SP2 and the corresponding partition of

the 4-tuple into pairs; (2) values Q1, Q2, Q3, R1, R2, R3; (3) partition of the 4-tuple into pairs corresponding to the "true" tree. We say that a 4-tuple is "good" if its partition based on the B-matrix of distances coincides with one based on the "true" tree. Otherwise the 4-tuple is called "bad".

Our experiments show that factor R1 is most adequate to distinguish good and bad 4-tuples, its value is in correspondence with quality of a phylogenetic tree based on the matrix. This is demonstrated by Table 1. The table shows the data for 3 genes giving the worst, the best and medium quality of NJ-tree (the quality of an algorithmic tree is characterized by the number of common edges between it and the "true" tree, see column NJ in the Table 1). At the next step of our work we plan to design a novel algorithm constructing phylogenetic tree using preliminary analysis of distant matrices.

NJ	NBad	MaxRBad	Good>0.3	Good>0.1		Bad>0.1	%%Bad>0.1
PF00542	12	8476	28.00%	1847	11333	1337	10.55%
PF01653	18	3595	19.00%	4033	15326	171	1.10%
PF01765	24	1984	14.00%	4230	16988	36	0.21%

Table 1. NJ – see the text; NBad – number of bad 4-tuples; MaxRBad – maximal value of R1 factor for bad 4-tuples; Good>0.3 - number of good 4-tuples with R-factor > 0.3 (no bad 4-tuples has such value of the factor); Good>0.1 and Bad>0.1 – analogous data for good and bad 4-tuples with the cut-off 0.1; %%Bad>0.1 – percent of bad 4-tuples among all 4-tuples with R-factor > 0.1

1. The Performance of Neighbor-Joining Methods of Phylogenetic Reconstruction. Algorithmica 25:251–278.

Structure Fluctuations and Configuration Instabilities in Proteins

Anatoly Ruvinsky[1], Ilya Vakser[1,2]

Center for Bioinformatics, The University of Kansas, USA ruvinsky@ku.edu
Department of Molecular Biosciences, The University of Kansas, USA

Structure fluctuations in proteins affect a broad range of cell phenomena, including stability of proteins and their fragments, allosteric transitions, and energy transfer. This study addresses thermodynamic and evolutional aspects of relationship between protein sequence composition, structure fluctuations and instabilities in proteins and their complexes. Structure fluctuations are characterized by a novel elastic network model accounting for the protein mass distribution and the interatomic interactions through a renormalized inter-residue Tirion-like potential.[1] We computed fluctuations for each of the protein residues in 184 proteins from 92 non-obligate protein-protein complexes selected from the DOCKGROUND non-redundant docking benchmark set (http://dockground.bioinformatics.ku.edu).[2] The results show that the residue mass and the structural environment determine the scale of the residue fluctuations. Surface residues undergo larger fluctuations than the core residues, in agreement with experimental observations.

We establish a new fluctuations-based classification of amino acids that includes three groups: (I) highly fluctuating - Gly, Ala, Ser, Pro, and Asp, (II) moderately fluctuating - Thr, Asn, Gln, Lys, Glu, Arg, Val, and Cys, and (III) weakly fluctuating - Ile, Leu, Met, Phe, Tyr, Trp, and His. The degree of hydrophilicity of groups I, II and III can be characterized as mixed, largely polar/hydrophilic and largely nonpolar/hydrophobic. The striking difference in the degree of hydrophilicity between residues in Groups II and III is explained by the non-homogeneous distribution of the amino acids in proteins. The polar residues prefer the surface and the nonpolar ones are more often located in the core. On the surface, a residue has fewer neighbors than in the core, and thus may be subject to greater fluctuations.

The clearly biased distribution of the groups I - III in proteins allows us to hypothesize that (a) the structural instabilities in proteins relate to a high content of the highly fluctuating residues and a lack of the weakly fluctuating residues in protein loops, chameleon sequences and disordered proteins, and (b) the nucleation of the unfolded phase starts from the protein loops and clusters of the highly fluctuating residues. The results point to strong correlation between the residue fluctuations and the sequence composition of the protein loops. This supports the

hypothesis on the origins of unfolding, and explains the inability of such sequences to form ordered secondary structure elements (α-helices or β-strands).

The above classification solves the longstanding contradiction of Vihinen *et al.* flexibility scale[3] that puts Gly, "generally considered to be the most flexible amino acid," in the middle of the flexibility scale and reveals limitations of using B-factors of the backbone only in describing protein flexibility. Our results show that Gly has the highest ability to fluctuate. Interestingly, the higher mobility of Gly has been commonly associated with the lack of the side chain, and thus greater conformational freedom. Our elastic network model operates with the network nodes, with each node replacing the corresponding residue. However, the mobile character of Gly is still well reproduced.

The results indicate that the fluctuations of the binding site residues, on average, are smaller than the fluctuations of the non-interface surface residues. The larger interface rigidity is primarily caused by Gly, Ala, Ser, Cys, Leu, and Trp. Formation of stable docking patches at the interface to facilitate the binding is discussed. The findings have broad implications for understanding thermostability of protein structures and their binding mechanisms, as well as mechanisms underlying the amino-acid propensities for irregular secondary structure elements (loops) and intrinsically disordered proteins.

1. A.M. Ruvinsky, I.A. Vakser (2011), Sequence composition and environment effects on residue fluctuations in protein structures, *J. Chem. Phys*., **133**: 155101.
2. Y. Gao et al. (2007), DOCKGROUND system of databases for protein recognition studies: Unbound structures for docking, *Proteins*, **69:** 845-851.
3. M. Vihinen et al. (1994), Accuracy of protein flexibility predictions, *Proteins*, **19**: 141-149.

Identification of date and party hubs in protein interaction network of *Saccharomyces Cerevisiae*

Mehdi Sadeghi[1], Babak Araabi[2], Mitra Mirzarezaee[3]

[1] *National Institute of Genetic Engineering and Biotechnology (NIGEB), Tehran, Iran*
[2], *School of Electrical and Computer Engineering, University of Tehran, Tehran, Iran*
[3] *Department of Computer Engineering, Islamic Azad University, Tehran, Iran*
sadeghi@nigeb.ac.ir

Proteins are important components of all living organisms. They are responsible for essential functions within cells. Most proteins perform their biological functions through interacting with other proteins [0]. Map of the whole physical protein interactions inside an organism forms a network called Protein Interaction Network (PIN). Although large-scale PINs have already been determined experimentally for several species; in general there is a lack of protein interaction data for many species and the computational prediction of protein interactions are still among the most wanted solutions in protein bioinformatics [**Error! Reference source not found.**]. These networks display scale-free topologies which are characterized by the power law distribution [0,0]. This means despite some negative remarks [0], in general a small fraction of proteins called hubs interact with many partners while majority of the proteins called non-hubs, interact with only a few others.

It has been understood that biological networks have modular organizations which are the sources of their observed complexity. Analysis of networks and motifs has shown that two types of hubs, party hubs and date hubs, are responsible for this complexity. Party hubs are local coordinators because of their high co-expressions with their partners, whereas date hubs display low co-expressions and are assumed as global connectors. However there is no mutual agreement on these concepts in related literature with different studies reporting their results on different data sets. We investigated whether there is a relation between the biological features of *Saccharomyces Cerevisiae*'s proteins and their roles as non-hubs, intermediately connected, party hubs, and date hubs. We propose a classifier that separates these four classes.

This study is focused on answering the following question: "*Which features should be used to better discriminate non-hubs, party hubs and date hubs in a PIN*?" A related sub-question is "*What classification methods more effectively discriminate these classes?*" In our experiments, we concentrate on *S. Cerevisiae's proteins* .

We extracted different biological characteristics including amino acid sequences, domain contents, repeated domains, functional categories, biological processes, cellular compartments, disordered regions, and position specific scoring matrix from various sources. Several classifiers are examined and the best feature-sets based on average correct classification rate and correlation

coefficients of the results are selected. We show that fusion of five feature-sets including domains, Position Specific Scoring Matrix-400, cellular compartments level one, and composition pairs with two and one gaps provide the best discrimination with an average correct classification rate of 77%.

This study also confirms the possibility of predicting non-hubs, party hubs and date hubs based on their biological features with acceptable accuracy. If such a hypothesis is correct for other species as well, similar methods can be applied to predict the roles of proteins in those species.

This work was supported in part by a grant from NIGEB (No.391) and IPM (No.CS1389-0-01).

1. A.L. Barabasi, Z.N. Oltvai (2004) Network biology: understanding the cell's functional organization. Nat. Rev. Genet. , **5**:101–113.
2. A. Tramontano A (2005) The Ten Most Wanted Solutions in Protein Bioinformatics. *Boca* Raton: Chapman & Hall/CRC
3. R. Albert (2005) Scale-free networks in cell biology. J Cell Sci, **118**: 4947–4957.
4. R. Tanaka et al (2005) Some protein interaction data do not exhibit power law statistics, FEBS Letters, **579**:5140-5144.

Reconstruction of Arabidopsis thaliana phosphatome

Dariya Samofalova, Pavel Karpov, Yaroslav Blume

Institute of Food Biotechnology and Genomics, Natl. Academy of Sci. of Ukraine, Kyiv, Ukraine,
samofalova_dariya@i.ua, : iht@i.kiev.ua

Protein phosphorylation plays a crucial role in biological functions and controls nearly every cellular processes, including metabolism, gene transcription and translation, cell-cycle progression, cytoskeletal re-arrangement, protein-protein interactions, protein stability, cell movement, and apoptosis. Protein phosphatases (PPs) play an important role in many signal transduction pathways, as the addition or removal of a phosphate group can activate or de-activate an enzyme, or allow a particular protein-protein interaction [1-3]. Protein phosphatases are divided into three large categories [4, 5]: serine/treonine-specific (protein phosphatase 1, 2A, 2B, 4, 5, 6, 7) and Mg2+/Mn2+-dependent protein phosphatases (PPM/PP2C), tyrosine-specific protein phosphatases and asparagin-specific protein phosphatases which differ by the presence of charachterising DXDXT/V catalytic domain motive [4, 6]. Unlike the protein kinases which have a common evolutional background, the groups of protein phosphatases evolved from different ancestral sequences. This fact predetermines their structural and functional differences [3, 4]. The aim of this investigation was the bioinformatic analysis of the plant phosphatome using Arabidopsis thaliana as a model.

138 types of proteins which contain protein phosphatase catalytic domain were identified on the basis of literature search and the analysis of UniProt database. Profile analysis (with using such instruments as SMART, PFAM and PROSITE) has shown that 105 protein phosphatases from A. thaliana contain serine/treonine-specific catalytic domain and 33 are tyrosine phosphatases. The database iterations were excluded based on gene coordinates information (according to Tair data).

Plant protein phosphatases were ranged on the basis of substrate specificity and structure of the catalytic domain and then divided into two groups. Group I was represented by serine/treonine-specific PP1, PP2A, PP2C, PP4, PP5, PP7, BSU, BSL families and phytochrome-associated PP. PP1 includes 8 representatives: P30366, P48482, P48483, P48484, P48485, O82733, O82734, Q9M9W3. Type PP2A is represented by 5 phosphatases: Q07099, P48578, O04951, Q8LAT9, Q8LAW8. The PP2C group has been the largest, numbering 80 protein phosphatases. PP4(PPX) includes 2 representatives: P48528 and P48529. Types PP5 (Q84XU2) and BSU (Q9LR78) are represented by single phosphatases. Group PP7 includes 3 phosphatases: Q9FN02, Q9LEV0, Q9LNG5. Type BSL is represented by 3 proteins: Q8L7U5, Q9SJF0, Q9SHS7. Phytochrome-associated PP1 (Q9SX52) and PP3 (Q9LHE7). It is shown that group I of protein phosphatases can contain catalytic domains of PP2Ac and PP2Cc types.

Group II has united the representatives of classical (Q84JU4, Q9ZVN4, Q8VZB2) and dual tyrosine phosphatases (Q8GY31). In addition it has been established that proteome A. thaliana contains 29 potential PTP (according to PTPc models and/or DSPc) which function is not yet discovered. To continue further reconstruction of structural homology and to prove functions of a number of PPs a search of optimal matrix PDB structures has been carried out.

Acknowledgments. The research was supported by STCU project grant #5215 “Search of effective protein phosphatase inhibitors using nanochemical approaches and evaluation of their biological activity in silico”

1. Arino J., Alexander D. Protein Phosphatases (Topics in Current Genetics). Springer. 2004. - 395p.
2. S. Luan (2003) Protein phosphatases in plants, Annu. Rev. Plant Biol, 54:63-92.
3. H. Wang et al. (2007) The protein phosphatases and protein kinases of Arabidopsis thaliana, In: Arabidopsis Book, Rockville, 1–38.
4. S. Almo et al. (2007) Structural genomics of protein phosphatases, J. Struct. Funct. Genomics, 8:121–140.
5. K. Wolstencroft et al. (2006) Protein classification using ontology classification, Bioinformatics, 14:e530–e538.
6. D. Kerk et al. (2008) Evolutionary radiation pattern of novel protein phosphatases revealed by analysis of protein data from the completely sequenced genomes of humans, green algae, and higher plants, Plant Physiol., 146:351–367.

Interaction between long and small ncRNAs

Nadine Albrecht, Hans-Werner Mewes, Thorsten Schmidt

Helmholtz Zentrum München – German Research Center for Environmental Health, Institute of Bioinformatics and Systems Biology (MIPS), Neuherberg, Germany, thorsten.schmidt@helmholtz-muenchen.de

Advances in transcriptomics and RNA-sequencing technologies pave the way for various research fields, including the reconstruction and identification of novel coding and non-coding transcripts [1]. Hence the amount of non-coding RNA of eukaryotic transcriptomes is still increasing by RNA-sequencing [2]. ncRNAs are recently classified according to their sequence length in small and long (>200 nt) ncRNAs, also referred as lincRNA [3,4]. These lincRNAs account for a large fraction of transcriptomes [3]. The current estimated number is about 17000 in the human and 10000 in the mouse genome [5]. They are capable to interact both, with transcripts or splice variants of protein-coding genes (pre-mRNA) and small ncRNAs [6]. This indicates that different coding and non-coding transcripts are not disconnected from each other, rather act jointly in a regulatory manner [6]. An interaction between a long and small ncRNA is associated with distinct functions [6]. For example, a few lincRNAs are reported to serve as precursor for small ncRNAs and to influence the generation of small ncRNAs (like endo siRNAs) [6]. Additionally some lincRNAs interfere the activity of small ncRNAs (like miRNAs) [6].

In this study, we determined the interplay of lincRNAs with distinct types of small ncRNAs using RNA-sequencing data including the reconstructions of novel lincRNAs across three mouse cell types: embryonic stem cells (ESC), mouse lung fibroblasts (MLF) and neural progenitor cells (NPC) [2], at large scale for the first time. To assess whether an interaction is existent, we run sequence similarity searches per tissue. Here we used Blat with significantly expressed lincRNA reconstructions (expression with p-value < 0.05 according to Guttman and colleagues [2]) and distinct types of small ncRNAs from public databases, DeepBase (http://deepbase.sysu.edu.cn/) and fRNAdb (http://www.ncrna.org/frnadb/). For all interaction sites, we calculated the expression levels as RPKM, according to Mortazavi et al [7] to identify expressed sites.

In this work, we found that lincRNAs interplay with a variety of small ncRNAs in all samples. 17% (947 of 5420) of lincRNA loci expressed in ESC, 13% (723 of 5437) in NPC and 8% (188 of 2295) in MLF overlap with a small RNA. LincRNAs interact especially with easRNAs, rasRNAs, pasRNAs, snoRNAs and miRNAs. For example, in the ESC sample 750 loci overlap sense with easRNAs, 109 with miRNAs of DeepBase and 128 with piRNAs, 175 with snoRNAs of fRNAdb.

This study was funded by Regulation and Evolution of Cellular Systems (RECESS).

1. B.J. Haas et al. (2010) Advancing RNA-Seq analysis, *Nat Biotechnol.*, **28:** 421-423.
2. M. Guttman et al. (2010) Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs, *Nat Biotechnol.*, **28:** 503-510.
3. P. Kapranov et al. (2007) RNA maps reveal new RNA classes and a possible function for pervasive transcription, *Science*, **316:** 1484-1488.
4. M.W. Wright et al. (2011) Naming 'junk': Human non-protein coding RNA (ncRNA) gene nomenclature, *Hum Genomics*, **5:** 90-98.
5. X. Wang et al. (2010) The Long Arm of Long Noncoding RNAs: Roles as Sensors Regulating Gene Transcriptional Programs, *Cold Spring Harb Perspect Biol*.
6. J.E. Wilusz et al. (2009) Long noncoding RNAs: functional surprises from the RNA world, *Genes Dev*, **23:** 1494-1504.
7. A. Mortazavi et al. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq, *Nat Methods*, **5:** 621-628.

Duplications of the neuropeptide receptor gene VIPR2 confer significant risk for schizophrenia.

Jonathan Sebat

UCSD, United States, jsebat@ucsd.edu

Rare copy number variants (CNVs) have a prominent role in the aetiology of schizophrenia and other neuropsychiatric disorders. Substantial risk for schizophrenia is conferred by large (>500-kilobase) CNVs at several loci, including microdeletions at 1q21.1 (ref. 2), 3q29 (ref. 3), 15q13.3 (ref. 2) and 22q11.2 (ref. 4) and microduplication at 16p11.2 (ref. 5). However, these CNVs collectively account for a small fraction (2-4%) of cases, and the relevant genes and neurobiological mechanisms are not well understood. Here we performed a large two-stage genome-wide scan of rare CNVs and report the significant association of copy number gains at chromosome 7q36.3 with schizophrenia. Microduplications with variable breakpoints occurred within a 362-kilobase region and were detected in 29 of 8,290 (0.35%) patients versus 2 of 7,431 (0.03%) controls in the combined sample. All duplications overlapped or were located within 89 kilobases upstream of the vasoactive intestinal peptide receptor gene VIPR2. VIPR2 transcription and cyclic-AMP signalling were significantly increased in cultured lymphocytes from patients with microduplications of 7q36.3. These findings implicate altered vasoactive intestinal peptide signalling in the pathogenesis of schizophrenia and indicate the VPAC2 receptor as a potential target for the development of new antipsychotic drugs.

NASP: A parallel program for identifying evolutionarily conserved nucleic acid secondary structures from sequence alignments

Jean Yves Semegni[1], Mark Wamalwa[2], Renaud Gaujoux[1], Gordon Harkins[2], Alistair Gray[1], Darren P.[1]

[1]*University of Cape Town, South Africa, jsemegni@gmail.com*
[2]*South Africa Bioinformatics Institute, South Africa*

Besides a capacity to store information within the sequences of their component nucleotides, single stranded nucleic acids can also potentially store information within their folded secondary structures. Under physiological conditions many single stranded RNA or DNA molecules nucleotides form meta-stable secondary structures which can have important roles in genome replication and gene expression. Although a number of computational methods exist for predicting nucleic acid secondary structures from either single sequences or alignments (Hamada et al., 2009; Markham and Zuker, 2008; Bernhart et al., 2008; Knudsen and Hein, 2003), even the best of these incorrectly infer a high proportion of base-pairings. Also, only a few methods provide any measures of statistical support either for their folding predictions, or for the over-all presence of secondary structure (Simmonds et al., 2004; Babak et al., 2007). From the perspective of experimental biologists seeking to test the functional relevance of secondary structures, it would be very useful to have a computational tool that, with the lowest possible false positive rate, will identify sites that pair within evolutionarily conserved secondary structures.

NASP (Nucleic Acid Structure Predictor) is an attempt to improve the selectivity with which individual secondary structures can be identified. It uses base-pairing probabilities provided by the UNAfold nucleic acid folding program hybrid-ss (Markham and Zuker, 2008).

The rationale behind NASP is simple: We assume that randomly shuffling nucleotides within sequences that have evolved to form stable secondary structures should influence their overall base-pairing potential such that the shuffled sequences should yield higher minimum free energy (MFE) estimates than the real sequences from which they were produced. By comparing MFE estimates made with real sequences to those made with randomized versions of these sequences, NASP tests whether there is evidence that the real sequences have greater structure forming capability than can be accounted for by chance. For each sequence, k, in an input alignment, hybrid-ss estimates the over-all Gibbs free energy of an optimally folded nucleotide sequence and yields a list of Boltzmann probabilities Pk(i, j) that nucleotide at position i and

nucleotide at position j form a base-pair. NASP then computes a consensus base-pairing matrix M, whose entry (i,j) is a weighted sum of the of the probabilities Pk(i, j).

NASP scans M through the anti-diagonal and recursively identifies groups of potentially base-paired nucleotides displaying the highest degree of evolutionary conservation (i.e. contiguous sets of non-zero entries along the anti-diagonal of M that have the highest sum)

Given such sequence alignments or even individual sequences as input, NASP recursively computes and outputs (1) the coordinates of potentially conserved stems and p-values indicating statistical support for additional unaccounted for secondary structures remaining in the sequences following each recursion, (2) the consensus structure in both the Vienna bracket-dot and a concatenation file formats and (3) the consensus base-pairing matrix, M, in both text and graphical formats.

We have tested NASP using known reference RNA structures and found that its over-all selectivity (the proportion of inferred base-pairs that are actually in the reference structures) is considerably better than that of Vienna's RNAalifold (Bernhart et al., 2008), Pfold (Knudsen and Hein, 2003), and CentroidFold (Hamada et al., 2009). The cost of NASP's low false positive rate is, however, a decrease in the true positive rate such that its over-all accuracy (measured here using the Mathews Correlation Coefficient, MCC, as described in Gardner et al., 2004) is slightly lower than that of Vienna (which was over-all the most accurate of the programs we tested). Nevertheless, we must stress that the primary focus of NASP is the identification of base-pairings with a false positive rate that is as low as possible: A focus that should prove particularly useful in studies aiming to evaluate the function of evolutionarily conserved (and therefore probably functional) nucleic acid secondary structures in that it should substantially reduce the time and expense needed to home in on those structures with the greatest biological relevance

Acknowledgements: We acknowledge the support of the Claude Leon Foundation

1.Babak,T. et al., (2007) Considerations in the identification of functional RNA structural elements in genomic alignments. BMC Bioinformatics, 8, 33.
2.Bernhart,S.H. et al. (2008) RNAalifold: Improved consensus structure prediction for RNA alignments. BMC Bioinformatics, 9:474.
3.Hamada,M. et al. (2009) Predictions of RNA secondary structure using generalized centroid estimators. Bioinformatics, 25:465-473.
4.Knudsen,B. and Hein,J. (2003) Pfold: RNA secondary structure prediction using stochastic context-free grammars. Nucleic Acids Research, 31 (13) 3423-3428.
5.Markham,N.R. and Zuker,M. (2008) UNAFold: software for nucleic acid folding and hybridization. Method Mol Bio., 453, 3-31
6.Simmonds,P. Tuplin,A. and Evans,D.J. (2004) Detection of genome-scale ordered RNA structure (GORS) in genomes of positive-stranded RNA viruses: Implications for evolution and host persistence. RNA, 10, 1337-1351.

Structure heterogeneity of particles of flexuous plant viruses

Pavel Semenyuk[1], Valentin Makarov[1], Anna Mukhamedzhanova[2], Evgeny Dobrov[1]

[1]*Moscow State University, Faculty of Bioengineering and Bioinformatics, Russian Federation,*
2 A.N. Belozersky Institute of Phisico-Chemical Biology, Russian Federation
psemenyuk@belozersky.msu.ru

Despite more than 50 years of studies high resolution structure information on any of flexuous helical plant viruses (FHPV) is still lacking. Best achievement in this direction is 14 Å date for potexvirus potato virus X (PVX) and potyvirus soybean mosaic virus, obtained by G. Stubbs laboratory with the help of fiber X-ray diffraction and cryo-electron microscopy [1]. Low resolution of fiber diffraction FHPV studies is explained by low degree of virions orientation. Potex- and potyviruses (and to a lesser degree – their isolated coat proteins (CP)) are also characterized by a whole number of different anomalies. In this work we present the results of far UV circular dichroism (CD) and analysis by different disorder prediction methods of some potexviruses and their CPs.

Until now the far UV CD spectra of only two potexviruses were reported. The Papaya mosaic virus (PapMV) spectrum, measured by Leclerc and co-authors contained no obvious anomalies and was similar to the spectrum of the isolated PapMV coat protein [2]. But the far UV CD spectrum of potato virus X virion measured 30 years earlier by Goodman had anomalous character and differed strongly from spectrum of isolated PVX CP [3]. In the present work we had measured far UV CD spectra for two more members of the Potexvirus genus: Alternanthera mosaic virus (AltMV) and Potato aucuba mosaic virus (PAMV) and their free CPs. The AltMV virion and AltMV CP spectra were similar to each other and to the spectra of PapMV and its CP. PAMV spectrum resembled PVX spectrum in anomalously low ellipticity of the negative band at 208 nm, but in contrast to PVX did not have additional peak at 228 nm.

The far UV CD spectra of the isolated potexvirus CPs were analysed with the help of K2D2 web-services and it has been found that this program gives for the potexvirus CPs (in contrast to the most of other proteins) the results disagreeing with the results of other methods. The potexvirus CP sequences were also analyzed by several unfolded regions prediction programs and these proteins have been found to probably contain significantly long disordered segments.

In the last 10 years PVX virions were found to undergo several structure transformations in the course of translation [4, 5] and on incubation at –20°C in low ionic strength solutions [6]. Therefore we suggest that FHPV virions may assume different conformations in different virions of the same preparation or even along the length of the same virus particle.

1. A. Kendall et al. (2008) Structure of flexible filamentous plant viruses. J Virol., 82: 9546 – 9554.
2. M.H. Tremblay et al. (2006) Effect of mutations K97A and E128A on RNA binding and self assembly of papaya mosaic Potexvirus coat protein. FEBS J., 273: 14–25.
3. R.B. Homer and R.M. Goodman (1975) Circular dichroism and fluorescence studies on potato virus X and its structural components. Biochim. Biophys. Acta, 378: 296 – 304.
4. J. Atabekov et al. (2007) Potato virus X: structure, disassembly and reconstitution. Mol. Plant Path., 8: 667 – 675.
5. E. Lukashina et al. (2009) Tritium planigraphy study of structural alterations in the coat protein of Potato virus X induced by binding of its triple gene block 1 protein to virions. FEBS J., 276: 7006 – 15.
6. M.A. Nemykh et al. (2008) One more probable structural transition in potato virus X virions and a revised model of the virus coat protein structure. Virology, 373: 61 – 71.

Small scale heterogeneity in mutation rate and mutation biases in Drosophila

Vladimir Seplyarskiy, Alexey Kondrashov, Georgii Bazykin

M. V. Lomonosov Moscow State University, Russia; Institute for Information Transmission Problems (Kharkevich Institute), Russian Academy of Sciences, Bolshoi Karetny pereu, Russian Federation, pamjat@mail.ru

Mutation rate is known to vary across the genome, depending both on nucleotide contexts and local chromatin properties. Here, we analyze the variation in the mutation rate and in the mutation bias, characterized by the transversion/transition ratio, across the genome of Drosophila melanogaster. First, by using data on triallelic SNPs within population of D. melanogaster, coincident biallelic SNPs between D. melanogaster and D. simulans, biallelic SNPs in D. melanogaster coinciding with divergence between D. yakuba and D. erecta, and sites of coincident D. melanogaster-D. simulans and D. yakuba-D. erecta divergence, we study nucleotide sites which underwent multiple mutations at the Drosophila phylogeny. In each of these analyses, we observe a significant excess, relative to the expectations, of positions with a pattern corresponding to multiple mutations. Furthermore, at sites of multiple mutations, there is a ~5% to ~35% increase in the transversion/transition ratio, relative to the genome average. Second, we study the local variation in the mutation rate and mutation biases along the genome of Drosophila. The transversion/transition ratio is significantly increased around the sites of transversions, but not around the sites of transitions, indicating local variation in the mutation biases at the scale of ~100 nucleotides. The transition/transversion ratio and the mutation rate vary dependently of each other on short scale, but this correlation totaly disappear at the distance more than 100 nucleotides. These results are not explainable by genome assembly errors, artefactual inference of ancestral states, biases in frequencies of ancestral polymorphism, or selection. They suggest a complex pattern of local variation in the mutation rates and the mutation biases along the genome. Possible mechanisms which could give rise to this pattern will be discussed.

Statistical Approach to Mutation Analysis of HIV-1 Primary Proteins

Roman Sergeev[1], Alexander Tuzikov[2], Vladimir Eremin[3]

[1]*Belarusian State University, 4,Nezavisimosti ave., 220030, Minsk, Belarus, roma.sergeev@gmail.com*
[2]*United Institute of Informatics Problems, Belarus*
[3]*Research Practical Center for Epidemiology and Microbiology, Belarus*

High variability of human immunodeficiency virus type 1 (HIV-1) is a major obstacle in its therapy. Mutations in specific parts of the virus genome may cause therapy failure because of drug resistance. Therefore, when choosing a therapy regimen one needs to know which mutations lead to drug resistance and how they relate to each other. The proposed approach allows to assess whether the identified mutations are associated with specific drugs as well as to determine their interdependence. This approach was practically implemented and used for preliminary analysis of sequences from HIV-infected patients from Belarus.

Polymerase pol gene of HIV-1 encodes two important virus proteins including protease (prot) and reverse transcriptase (RT). Most drugs are used to inhibit one or another protein and specific inhibitor mutations may occur during therapy in regions coding both enzymes [1]. This might lead to the fact that such mutated copies of the virus become insensitive (resistant) to the inhibitor. These mutations are well documented for HIV-1 subtype B [2]. Secondary mutations can also occur at the same time, including compensatory mutations that have been little studied. In this connection it is necessary to develop and implement an approach to study correlations between mutations in a chosen position with mutations in all other positions as well as to describe possible additional resistance mutations in HIV-1 non-B subtypes [3].

To analyze a possible relation between the identified mutations and antiretroviral therapy we apply contingency tables. A table of occurrences is constructed to determine whether a mutation in a specified position of the sequence is caused by some drug. Chi-square criteria or Fisher exact test followed by multiple hypothesis testing are used to make conclusions about associations.

To find coevolving sites we use several approaches. Firstly, we construct a phylogenetic tree for the alignment and perform ancestral sequence prediction to get sequences in its internal nodes. Amino acid substitutions are mapped into its branches to get mutation history of each site. Obtained mutation histories are linked with each site and used in further analysis. We also exclude some non-informative sites from consideration. There are sites which based on their mutation histories are likely to be conservative and sites with relatively small values of mutual information. The associations between remained sites are evaluated using correlation coefficients

computed under obtained mutation histories for each pair of sites.

To understand the contributions of mutations in establishment of drug resistance we apply the linear regression. Mutations in particular positions form independent variables while virus load was quantized into three levels being a dependent variable.

All calculations related to the analysis of primary protein sequences were performed using third-party open source software (ClustalW for multiple sequence alignment, RAxML for phylogenetic analysis, GASP for the ancestral prediction) and own utilities developed in C++ and R languages. Experimental data for the analysis were provided by the laboratory of diagnosis of HIV and opportunistic infections of the Republican Research and Practical Center for Epidemiology and Microbiology (Minsk, Belarus) as a set of gene pol sequences including patients' charts.

1. R.W. Shafer et al. (2007) HIV-1 protease and reverse transcriptase mutations for drug resistance surveillance, *AIDS*, **21(2):** 215-223.
2. Stanford University HIV Drug Resistance Database [Electronic resource]. – 1998-2011. – Mode of access: http://hivdb.stanford.edu/
3. M.S. Hirsch et al. (2008) Antiretroviral Drug Resistance Testing in Adult HIV-1 Infection: 2008 Recommendations of an International AIDS Society–USA Panel, *HIV/AIDS CID*, **47:** 266-285.

Bioinformatic Analysis of Structural Factors of Selective Inhibition in Human Protein Kinase C Family

Daria Shalaeva[1], Vakeel Takhaveyev[1], Dmitry Suplatov[2], Vytas Švedas*[2,1]

[1]*Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University, 119991, Russia, Moscow, Vorobjev hills, 1-73, Russian Federation, shalaeva@belozersky.msu.ru*
[2]*Belozersky Institute of Physicochemical Biology, Lomonosov Moscow State University, 119991, Russia*
* Corresponding author - *vytas@belozersky.msu.ru*

PKC isoforms are involved in transmembrane signal transduction pathways, regulating a variety of cellular functions such as growth, differentiation, tumor promotion and apoptosis. Individual PKC isoforms differ in their expression patterns and substrate specificities, strongly suggesting that each isoform may be involved in distinct regulatory processes within the cell. Activation of PKC isoforms occurs in a number of pathological states. Availability of isoform-selective PKC inhibitors may provide important pharmacological agents to better define the physiological and pathological functions of each isoform.

Comparative bioinformatics analysis of homologous PKC enzymes was used to study molecular determinants of selective inhibition. Subfamily specific positions (SSP) – variable amino acid residues with a tendency to be conserved only within a subfamily of enzymes, but different between subfamilies - are responsible for functional divergence within families of

homologous enzymes. A library of human PKC catalytic domain structures was created using files from PDB for PKCα, PKCβ and PKCθ and homology modeling to predict three-dimensional representations of PKCγ, PKCδ, PKCε and PKCη. Multiple sequence alignment of 89 PKC enzymes from different species was created. Both conserved and specific positions in PKC family of enzymes were identified and further explored using molecular docking.

Molecular docking and molecular dynamic simulation of PKC complexes with known inhibitors LY333531, Gö6978, GF109203x, Ro 31-8220 were used to train a set of geometry filters of productive binding of the inhibitor in different isoforms based on its orientation to known catalytically important conserved residues Val-423, Glu-421 and Thr-404 (in PKCβ).

In silico library of 54 new potential selective inhibitors was generated to study the role of subfamily-specific positions in specific binding among different PKC isoforms using molecular docking. Ala-483 in PKCα and PKCβ was found to be the most promising specific position due to evident stereo chemical differences with homologous threonine in PKCγ. Molecular docking shows that introduction of methylamine group instead of indole ring in Gö6978 inhibitor disables the binding to PKCγ isoform (Fig.1).

Thus, bioinformatics analysis and molecular modeling reveal subfamily-specific position responsible for discrimination of inhibitor binding to ATP-binding site of human protein kinases and outline the new computational strategy in modeling selective inhibitors of human PKC enzymes.

This work was supported by Russian Foundation of Basic Research grant 09-04-92744-ННИОМ_а

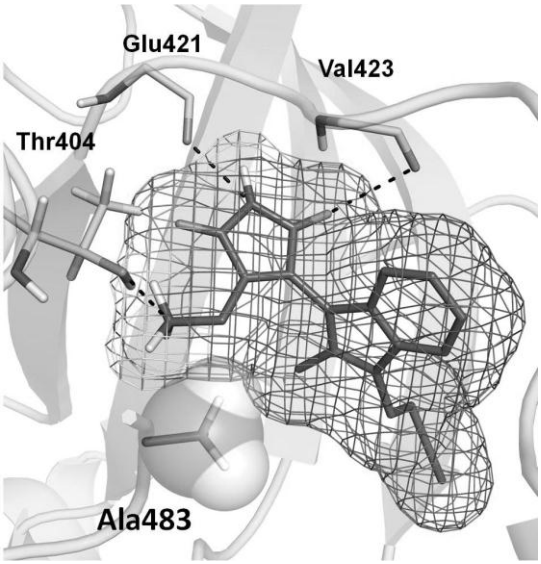


Fig.1. Complex of PKCβ with modified Gö6978 inhibitor. Introduction of methylamine group instead of indole ring in Gö6978 inhibitor disables the binding to PKCγ that has Thr instead of Ala483.

Molecular phylogeny of the genus Silene L. Section Auriculatae (Caryophyllaceae)

MASOUD SHEIDAEI

Shahid Beheshti University, GC., Faculty of Biological Sciences, Tehran, Iran,, Iran, msheidai@yahoo.com

Morphological and RAPD studies were performed on Silene species of the sect. Auriculatae growing in Iran by using phenetic, parsimony and Bayesian analyses. Trees obtained differed in the species groupings although agreed in some parts. Parsimony and Bayesian analyses of morphological characters produced some clades which were not well supported by bootstrap and clade credibility values but UPGMA tree showed a high Cophenetic correlation. Grouping based on morphological characters partly support the species affinity given in Flora Iranica. Out of 40 RAPD primers used 15 primer produced reproducible polymorphic bands. In total 347 bands were produced out of which 340 bands were polymorph and 7 bands were monomorph. Among the species studied S. goniocaula showed the highest number of RAPD bands (184), while S. commelinifolia var. isophylla showed the lowest number (123). Some of the species studied showed the presence of specific bands which may be use for species discrimination. NJ and Bayesian trees of RAPD data partly agree with morphological trees obtained.

Molecular dynamics of prototype foamy virus protease flap region, N- and C-termini in aqueous solution

Sergey Shityakov, Thomas Dandekar

[1]*Würzburg University, Germany, Shityakov@vim.uni-wuerzburg.de*

Molecular dynamics trajectories are Cartesian coordinates produced by recording of dynamical changes over time representing the positions of each atom along a series of small time step. Here, we implemented atomistic molecular dynamics simulations of prototype foamy virus (PFV) protease monomer to investigate the conformational changes of the flap region, N- and C-termini in aqueous solution. The PFV protease monomer undergoes some changes of secondary structure but remains stable during 10 ns simulation time. In particular, the flap region and the N- and C-termini turned out to be highly flexible. Nevertheless, retroviral protease dimerization process occurs through the anti-parallel β-sheet, which is absent in the PFV protease. Although the overall folds of β-sheets and α-helices are remained quite similar and stable, the PFV protease dimerization mechanism reveals significant differences in the dimerization interface relative to other retroviral proteases, such as HIV protease. Therefore, PFV protease dimerization event might be mediated through the additional viral or cellular cofactors. Finally,

the results provide a model for the flap region, N- and C-termini overall dynamics that is considered to be important for regulation of the enzyme function.

SimZoom: An Exploration Environment for Coalescent Simulation Traces

Ilya Shlyakhter, Pardis Sabeti

Harvard University, Cambridge, MA, USA, ilya_shl@alum.mit.edu

When coalescent simulators generate a simulation, they produce, in addition to the genotypes of the samples from each population, an Ancestral Recombination Graph [2] representing the complete history of the sample. While there are many tools for analyzing the resulting sample (same tools as for analyzing real population genetics data), there are few tools for interactively exploring the ARG. In real data, the ARG is unknowable, though tools exist for inferring plausible ARGs [2]. But for simulations, analyzing the ARG can help understand why a particular statistical method -- for example, a statistic that discriminates between neutral SNPs and SNPs under selection -- did not work well in a particular simulation. Such statistics are defined to be computed on the present-day sample of genotypes, but the intuition behind them is often formulated in terms of the history of the sample. For example, long-range haplotype tests [3] use haplotype breakdown of the original haplotype on which an allele is born as a proxy of the allele's age: for a recent allele, most chromosomes carrying the allele will also carry a large chunk of the original mutation's neighborhod; for an older allele, the chunk of original neighborhood carried on most chromosomes will be much shorter. Occasionally, this proxy for age fails, giving an abnormally broken-down haplotype for a recent allele or an abnormally preserved haplotype for an older one. SimZoom can be used to investigate and classify the causes of such anomalies.

An Ancestral Configuration (AC) is a set of lineages that existed at a particular past generation [2]. Each lineage carries defined chromosome segments inherited by at least one present-day chromosome; the remaining chromosome portions are not represented. SimZoom's main display shows the ancestral configuration for a particular generation, as a table where each row represents one haplotype and each column represents one SNP. It allows the user to move forward and backward in time by specified intervals, observing the changes in the ancestral sample. More important, it allows the user to mark a subset of chromosomes and a sub-region on each chromosome in the subset, and then track the ancestors/descendants of these regions as the view moves to other generations. Such markers are termed "hourglasses", as they tend to denote an hourglass-shaped portion of the ARG (narrow at the generation of definition, expanding as you move forward/backward in time from that generation). Multiple hourglasses can be defined at the same time, represented in the view by different colors. By combining multiple hourglasses, it becomes simple to track the ancestry of particular groups of lineages and/or particular regions of the simulated chromosome. Hourglasses can be defined by manually selecting the relevant portions of the relevant lineages, but also by automatically selecting SNPs and/or chromosomes

meeting specified conditions expressed in terms of tracks (described below). There are also custom commands to define hourglasses: for example, to include in an hourglass all chromosome segments identical by descent to a chosen segment.

Time navigation commands can be restricted to time points relevant to a particular hourglass. While the full ARG is typically too complex to navigate interactively, the number of events relevant to a given hourglass -- time points at which the set of lineages in the hourglass undergoes some change -- is typically small enough to step through. The user can choose what types of events -- coalescences/recombinations, mutations and migrations -- to consider when finding the next relevant event in a given time direction. There is also a command to move in a given time direction until a specified condition becomes satisfied.

SimZoom provides a highly customizable and extensible mechanism for visualizing ancestral configuration informations, based on the notions of *tracks* and *renderers*. A track computes a value for each SNP, for each chromosome, each SNP/chromosome pair, or one value for the entire ancestral configuration. How values are plotted is controlled by renderers; there are several predefined renderers for common value types, and new renderers can be added easily as plug-ins. Track values are plotted above or to the side of the view, for per-SNP or per-chromosome tracks respectively. Tracks that give a value for SNP/chromosome pairs are mapped by renderers to attributes of main view cells, so that a mix of several values can be plotted: for example, cells that belong to several hourglasses can be colored with a mix of the corresponding colors. Track values can be used to filter the display, limiting it to SNPs and/or chromosomes that meet specified conditions.

Tracks are defined as Java expression strings that can reference SNP and chromosome attributes, other tracks, and many auxiliary functions. Many predefined tracks are available, and new ones can be by adding/modifying the expression strings. New expression building blocks can be defined by adding simple plug-ins implemented in Java. This architecture makes SimZoom an extensible architecture for visualizing and experimenting with ARG-related algorithms, in the same way that CytoScape [4] is an extensible architecture for network algorithms.

There is a command to plot the values of particular tracks over time. This allows viewing the variance over time of, for example, the frequencies of particular SNPs in given populations, or of linkage disequilibrium between specified SNPs. Whenever possible, time points at which the given track values may change are automatically determined, leading to fast plotting of track values across large time ranges. The plots are hyperlinked to the main view, so that clicking on a particular generation on the plot switches the main view to that generation.

The current version works with a version of the cosi coalescent simulator [1], modified to output ARG; future versions will support import from other simulators.

1. S. Schaffner et al. (2005) Calibrating a coalescent simulation of human genome sequence variation, Genome Research, 15:1576–1583.
2. M.J.Minichello and R.Durbin (2006) Mapping Trait Loci by Use of Inferred Ancestral Recombination Graphs, American Journal of Human Genetics, 79(5):910–922.
3. P.Sabeti et al. (2002) Detecting recent positive selection in the human genome from haplotype structure, Nature, 419:832–837.
4. M. Smoot et al. (2011) Cytoscape 2.8: new features for data integration and network visualization, Bioinformatics, 27(3):431–432.

Visualization of the MS-Align algorithm results for the protein spectrum matches

Yakov Sirotkin[1], Xiaowen Liu[2], Yufeng Shen[3], Gordon Anderson[3], Yihsuan S. Tsai[4], Ying S. Ting[4], David R. Goodlett[4], Richard D. Smith[3], Vineet Bafna[5] and Pavel A. Pevzner[5]

[1]*St Petersburg Academic University, Russian Federation, yasha@telamon.ru*
[2]*University of California, San Diego, United States*
[3]*Biological Science Division, Pacific Northwest National Laboratory, United States*
[4]*Department of Medicinal Chemistry, University of Washington, United States*
[5] *Department of Computer Science and Engineering, University of California, San Diego, United States*

The results of the protein identification using top-down spectra produces a lot of data that are hard to show and understand. We post-process the results and generates standard html-files. It provides lighting-fast interface for the scientists and allows them to share their findings with colleagues.

Protein is Coded in Genome and Synthesized in Ribosomes as a Structural Template of a Rotameric Version Sequence of Peptide Bound Configuration

Victoria Sokolik

SI «Institute of neurology, psychiatry and narcology of the AMS Ukraine», Ukraine, sokolik67@rambler.ru

To date, algorithms of modeling spatial structure of protein start with its amino asid sequence. Everything began with Anfinsen C.B., which was the first to show the presence of interrelation between amino asid sequences and a protein conformation. Consideration of peptide bounds between amino asids before involuntarily escapes the concept of primary structure of protein while discuss its amino asid sequences. We have simulated three rotameric configurations of peptide bounds as basic elements of a structural template of protein, determined by the third nucleotide of each codon. For experimental substantiation of the given idea analysis of frequency of occurrence amino asid determined by codons with certain nucleotides in the third position of codons in the secondary structure of protein conformers (helix and strand) has been carried out in a random collection of 35 proteins (Protein Data Bank – PDB). The collection included functionally various proteins: apolipoproteins, cytochromes, enzymes, histones, hemoglobin, albumin, interleukin, interferons and so on, for which the scheme of the secondary structure was full of PDB and nucleotide sequences, were available in EMBL.

The analysis of frequency occurrences (in %) XYA codons, ≥3 successive codons XYU or XYC/G (determine the minimum coil of the helix from 3 amino asids) in nucleotide

sequences, coding experimentally revealed helix and strand in these proteins, showed a tendency to prevail coding of the helix by XYC/G codons for all the pool of the examined proteins. Further we supposed that the posttranslational folding in different degree contributes in updating of the coded individual structural template of proteins. The search of proteins with the biggest level of determination in genome of the secondary structure were made by the way of comparison the secondary structure schemes of PDB (made according to experimental data) with corresponding schemes, decoded according to nucleotide sequences (in program Secondary Structure Protein – SSP). Nine such proteins (*Apolipoprotein C-I, Cytochrome P450 1A2, Cathepsin B, Histone H4, Histone H2A, Gonadotrophin alpha chain, Gamma-synuclein, Hemoglobin subunit alpha, Somatotropin*) was revealed in all the pool of the examined proteins. Folding doesn't influence essentially their conformation; therefore we managed to establish the specific coding of the helix by XYC/G codons and strands by XYA codons (Fig.1). For helixes were not differentiated on the right and left in PDB schemes, it's a pity. We have assumed that the given proteins are evolutionary ancient and appeared before the system of chaperones. The diagram of distribution of XYC/G/A/U codons in conformers of the secondary structure of protein structural templates, decoded in the SSP program, is presented in Figure 2. The specific coding of protein structural template was managed to distinguish owing to minimization of folding noise, which was ignored by other researchers, who didn't discovered any law even at the analysis huge protein pools. The way of coding of peptide bound rotamerism in the genome and a mechanism of decoding of structural templates of proteins in ribosome are discussed.

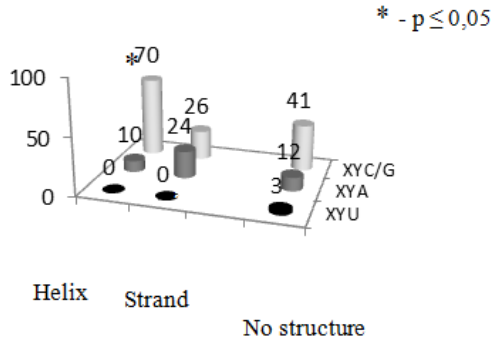


Fig. 1. Nine proteins (PDB data): the frequency of occurrences (in %) of XYA, XYU and XYC/G in nucleotide sequences, coding helix and strand

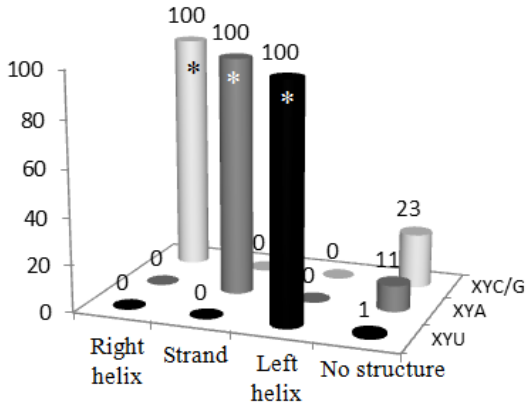


Fig. 2. All protein (SSP data): the frequency of occurrences (in %) of XYA, XYU and XYC/G in nucleotide sequences, coding right or left helix and strand

Thus, proteins are synthesized in a ribosome in the form of an individual structural template with secondary structure determined in genome, but not just as a kind of unstructured amino asid sequences. A native conformation of protein synthesized de novo gets as a result of updating of a structural template during the posttranslational folding. These ideas should become a basis of technology of matrix synthesis of new synthetic proteins with structural templates which can be coded under the set functions.

Vertical evolution and horizontal transmission of Tc1/ *mariner* superfamily DNA transposons in Lepidopteran species

I. Sormacheva, A. Novikov, and A. Blinov

Institute of Cytology and Genetics, Pr. Lavrent'eva 10, Novosibirsk-90, Russian Federation, 630090, sormacheva@bionet.nsc.ru

Mobile elements (ME) are able to change their location in the genome and play an important role in the organization, functioning and evolution of genome. ME from ***Tc1/mariner* superfamily** are probably the most diverse and widespread elements of class DNA transposons in nature. According to the modern classification the ***Tc1/mariner*** superfamily is devided into three families: *mariner*, *ludens*, and *mori*. The *mariner* family includes fifteen subfamilies: DTTMar (Mariner), DTTMarAtl (atlantis), DTTMarBRI (briggsae), DTTMarCAP (capitata), DTTMarCec (cecropia), DTTMarCRI, DTTMarELE (elegans), DTTMarGGS, DTTMarIRR (irritans), DTTMarLin (lineata), DTTMarMau (mauritiana), DTTMarMel (mellifera), DTTMarROS (rosa), DTTMarUrt, DTTMarVer (vertumnana). Two subfamilies DTTLudGAN and DTTLudCAE belong to the *ludens* family, and only one subfamily DTTMOR belongs to the *mori* family. Vertical or horizontal (cross-species) transfers are postulated as two alternative ways in which DNA transposons evolve. ***Tc1/mariner*** elements can be horizontally transferred between reproductively isolated species.

To investigate the evolution of ***Tc1/mariner*** elements we studied a distribution of the representatives of this superfamily in the genomes of different species from the order Lepidoptera, using both bioinformatical and experimental approaches. Bioinformatic approach. At this moment seven ***Tc1/mariner*** elements have been described for the genome *Bombyx mori* (the only one completely sequenced genome among Lepidoptera species): Bmmar3 and Bmmar5 (DTTMarCec subfamily), Bmmar2 (DTTMarCRI subfamily), Bmmar4 (DTTMarMel subfamily), Bmmar1 and Bmmar6 (DTTMOR subfamily), and BmmarY (DTTMarVer subfamily). We perform search across databases to find all available ***Tc1/mariner*** elements from the different insect species. In addition to the Bmmar elements from *Bombyx mori*, 72 nucleotide sequences of ***Tc1/mariner*** elements have been isolated from the GenBank database using nucleotide blast search. The phylogenetic analysis demonstrated a presence of both the *mori* and *ludens* families, and ten subfamilies of the *mariner* family in the insect genomes. ***Tc1/mariner*** elements from the Lepidoptera species have been found in six subfamilies of the

mariner family (DTTMarCec, DTTMarMel, DTTMarVer, DTTMarMau, DTTMarIrr, and DTTMarCRI) and in the *mori* family (DTTMOR).

Experimental approach. Total DNA from 20 lepidopterian species were screened by PCR reaction using degenerate primers MAR-124F and MAR-276R specific to transposase gene of the ***Tc1****/mariner* elements. The PCR products were isolated, cloned in to the plasmid vector, and sequenced. Subsequent comparison with the sequences of ***Tc1/mariner*** elements available from databases and phylogenetic analysis showed that three (DTTMarCec, DTTMarMel, DTTMarMau) of seven ***Tc1/mariner*** groups described in butterflies, were detected using PCR amplification with degenerative primers. The similarity level of DTTMarCec, DTTMarMel, DTTMarMau sequences and distribution within lepidopterian species showed its vertical evolution. Further we have focused our investigation on the rest three groups of ***Tc1****/mariner* elements: DTTMOR, DTTMarVer, and DTTMarCRI, represented in the genome *Bombyx mori*, but not identified by PCR amplification with degenerate primers in the other species. We made a PCR search of these elements with the primers specific for each of these groups, designed on sequences of the known Bmmar transposons from *Bombyx mori*. Results of PCR amplification showed *uneven distribution* of two elements (BmmarY and Bmmar1)among studied *Lepidoptera* species. The distribution of Bmmar1 (DTTMOR family) and BmmarY (DTTMarVer subfamily) elements in butterflies is bordered by two genera - *Maculinea* and *Bombyx* that lead us to a suggestion of the horizontal transfer DTTMarVer and DTTMOR elements between *Maculinea* and *Bombyx* species. Phylogenetic analysis showed a high similarity between elements from the evolutionary distinct species. The value of BmmarY elements divergence between two genera (3,4%) is comparable with intraspecies divergences in these species (*Maculinea* – 2,8 % and *Bombyx* genera – 2,5%, correspondingly) . The average intraspecies divergence between *Maculinea* and *Bombyx* species is 6,6% (interspecies nucleotide divergence for the *Maculinea* and *Bombyx* species - 7,0% and 4,0%, correspondingly). The high phylogenetic similarity between ***Tc1****/mariner* elements from evolutionary distinct species confirms our suggestion of the horizontal transfer of DTTMarVer and DTTMOR elements between *Maculinea* and *Bombyx* species.

Comparison of phylogeny reconstruction programs on sequences of fungal protein domains

Mikhail Krivozubov[1], Sergei Spirin[2]

[1]*Faculty of Bioengineering and Bioinformatics of Moscow State University, Russian Federation*
[2]*Belozersky Institute of Moscow State University, Russian Federation, sas@belozersky.msu.ru*

We compared efficiency of eight popular free programs of reconstruction of protein phylogeny. Those programs are: PhyML [1] (versions 2 and 3), FastME [2], FastTree [3], and four programs from the PHYLIP package [4], namely, protpars, proml, fitch and neighbor. All programs are tested with the options that are set by default.

The material for comparison was protein sequences of 26 species of Fungi. The criteria for selecting the species were as follows: first, in Pfam database there are at least 2000 sequences of evolutionary protein domains from each species; second, we succeeded in obtaining sequences of 18S and 25S ribosomal RNAs of the species.

From Pfam database [5], we extracted sets of protein sequences called "series of probably orthological protein domains" (SPOD). The criteria for regarding a set of sequences as a SPOD are as follows. First, the set consists of sequences from one Pfam family. Second, each of 26 species should be represented in the set by exactly one sequence. Third, for each pair of species, A and B, the sequence representing A should be the closest relative for the sequence representing B, among all sequences of Pfam domains from A (and vice versa).

We can expect that protein domains of a given SPOD are real orthologs and thus their phylogeny coincides with the phylogeny of the species. That may be not right for some SPODs (because of possible loss of paralogs in sequence databases or in real proteoms). The latter adds a "noise" to our data but should not completely hide the "signal".

With the program Muscle [6], we created an alignment of each SPOD. From our data, we removed all alignments containing less than 20 informative (non-conserved and non-gap) columns; after that, 1864 alignments remain. Using each of the remaining alignments, we constructed eight phylogenetic trees with eight programs. Also we constructed two "model" trees of the species, using sequences of 18S rRNA and 25S rRNA.

For two tested programs, P and S, and an alignment X, we regarded P as working better than S on X if the tree created by P is closer to both model trees. Concerning the entire set of alignments, we regarded P as working better than S if the number of alignments for which P worked better was greater than the number of alignments for which S worked better, and the difference was statistically significant.

We obtained no significant difference between the results of six programs: PhyML (both versions), FastTree, FastME, neighbor, and fitch. The program proml (the realization of the maximum likelihood algorithm in PHYLIP) worked significantly worse than each of the programs mentioned above; and the program protpars (the realization of the maximum parsimony algorithm in PHYLIP) worked significantly worse than all other programs.

Our work was partly supported by the joint grant of Russian Foundation for Basic Research and DFG (grant number 09-04-92743).

1. Guindon S., Gascuel O. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Systematic Biology, 52: 696-704.
2. Desper R., Gascuel O. (2002) Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. Journal of Computation Biology, 9: 687–705.
3. Price M.N., Dehal P.S., Arkin A.P. (2009) FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol. Biol Evol., 26: 1641-1650.
4. Felsenstein, J. (1989) PHYLIP – Phylogeny Inference Package (Version 3.2). Cladistics, 5: 164-166.
5. Finn R.D, et al. (2010) The Pfam protein families database. Nucleic Acids Research, 38: Database Issue D211-D222.
6. Edgar R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research, 32: 1792-1797.

Genomic analysis of transcriptional regulation of aromatic amino acid metabolism in gamma-proteobacteria

Vita Stepanova[1], Dmitry Rodionov[2]

[1]*Institute for Information Transmission Problems RAS, Russian Federation, vita.stepanova@gmail.com*
[2]*Sanford-Burnham Medical Research Institute, United States*

TyrR and TrpR transcription factors play a crucial role in the aromatic amino acids metabolism. Here we apply a comparative genomics approach to analyze TyrR and TrpR regulons in the genomes of gamma-proteobacteria. TyrR regulon was reconstructed in the Enterobacteriales, Alteromonadales, Vibrionales, Pseudomonadalles, Pasteurellales, and Aeromonadales groups. It was shown that the regulon content varies significantly between groups of proteobacteria. The positive or negative mode of TyrR action was predicted. TrpR regulons was reconstructed in the Enterobacteriales, Alteromonadales, Vibrionales, Pasteurellales, Xanthomonadales, Psychromonadaceae, Oceanospirillales, Moraxellaceae and Shewanellaceae groups of proteobacteria. Multiple novel members of both TyrR and TrpR regulons were discovered. All reconstructed regulons are available for browsing in the RegPrecise database within the TyrR and TrpR collections (http://regprecise.lbl.gov/).

Towards the molecular archtecture of intermediate filaments

Sergei Strelkov

Katholieke Universiteit Leuven, Belgium, sergei.strelkov@pharm.kuleuven.be

Alongside with microtubules and actin filaments, intermediate filaments (IFs) form the third principal filament system contributing to animal cytoskeleton [1]. IFs perform an essential structural function within many cell types, and it is not surprising that a fast growing number of mutations in IF proteins has been associated with currently incurable human diseases such as myopathies, skin and neuronal diseases and even premature ageing. At the same time, the molecular architecture of IF remains an elusive subject despite of many years of research, in stark contrast with the well-established architectures of microtubules and F-actin [2]. We aim at gaining the understanding of the IF architecture and assembly process via a bottom-up approach, i.e. starting with the atomic structure of the elongated IF dimer, which serves as the elementary building unit of the filament. We use a 'divide-and-conquer' strategy based on X-ray crystallography of multiple fragments, a method that allowed us to gradually resolve the nearly complete coiled-coil dimer of human IF protein vimentin [3]. Furthermore, we use modelling approaches to create the 3D structures of tetramers, higher assembly intermediate and finally complete filaments. Here, electron microscopy and small-angle X-ray scattering provide important restraints that guide and verify *in silico* approaches. Finally, comparative primary structure analysis of various IF proteins together with the growing knowledge of their 3D organization allow us to make conclusions about the characteristic 'signature features' of IF proteins. Recently, bacterial protein crescentin and insect protein isomin were proposed to present new classes within the IF family. However, we argue that these propositions require further proof that should be based on the detection of IF signature features in these novel proteins [4].

1. Herrmann H, Bär H, Kreplak L, Strelkov SV, Aebi U (2007) Intermediate filaments: from cell architecture to nanomechanics, *Nature Reviews Molecular Cell Biology* **8**:562-573.
2. Strelkov SV, Herrmann H, Aebi U (2003) Molecular architecture of intermediate filaments, *Bioessays* **25**: 243-251.
3. Strelkov SV, Herrmann H, Geisler N, Lustig A, Ivaninskii S, Zimbelmann R, Burkhard P, Aebi U (2001) Divide-and-conquer crystallographic approach towards an atomic structure of intermediate filaments, *Journal of Molecular Biology*, **306:** 773-781.
4. Herrmann H, Strelkov SV (2011) History and phylogeny of intermediate filaments: Now in insects, *BMC Biology*, **9:** 16.

TSAR — a new graph-theoretical approach to computational modeling of protein side-chain flexibility.

Oleg Stroganov[1], Fedor Novikov[1], Alexey Zeifman[2], Viktor Stroylov[1], Ghermes Chilov[1]

[1]*MolTech Ltd, Russian Federation, ostroganov@moltech.ru*
[2]*N.D.Zelinsky Institute of Organic Chemistry, Russian Federation*

A new graph-theoretical approach called TSAR (Thermodynamic Sampling of Amino acid Residues) has been elaborated to account for the protein side chain flexibility in modeling conformation-dependent protein properties. In TSAR, a protein is viewed as a graph whose nodes correspond to structurally independent groups and whose edges connect the interacting groups. Each node is assigned a set of conformational and ionization states of the corresponding group, and each edge is assigned an array of interaction potentials between the adjacent groups. The partition functions of each node are found by treating the obtained graph as a belief-network - a well-established mathematical object.

TSAR was applied to assess ionization properties of protein from explicitly calculated partition functions of the ionized forms of protein residues. A simplified version of a molecular mechanical scoring function, borrowed from the Lead Finder docking software, was used for energy calculations. The accuracy of the resulting model was validated on a set of 484 experimentally determined pKa values of protein residues. The average correlation coefficient between calculated and experimental pKa values was 0.80, ranging from 0.95 (for Tyr) to 0.61 (for Lys). It appeared that the contribution of electrostatic interactions to the pKa values was relatively small compared hydrogen bond energy contribution, suggesting that the balance between short- and long-range interactions in determining the protein ionization properties is yet to be refined.

The effectiveness of the TSAR approach for sampling of the protein conformational space was demonstrated by benchmark for protein side-chain prediction problem. TSAR algorithm together with dead-end elimination techniques was used to calculate coordinates of protein side-chains given position of its backbone. The algorithm was able to complete predictions on a set of 65 proteins with total chi(1) and chi(1 + 2) dihedral angle accuracies of 88% and 79% using a backbone-dependent rotamer library and simplified version of a molecular mechanical scoring function.

Bayesian Inference of Protein Complexes from Mass Spectrometry Data

Alexey Stukalov, Jacques Colinge

Research Center for Molecular Medicine, 1090 Vienna, Austria, astukalov@cemm.oeaw.ac.at

Affinity purification assays combined with Mass Spectrometry (AP/MS) provide a powerful experimental technique for the discovery of protein-protein interactions and protein complexes. We propose a novel statistical method to infer protein complexes from initial output of AP/MS experiments, which may contain many non-specific proteins and lack some true interactors.

While for the analysis of large-scale AP/MS datasets (thousands of bait proteins), one may use methods based on the detection of highly connected network components [1,2], they are not suitable for small- and medium-scale datasets (10-100 baits). For such datasets, methods based on semi-quantitative MS data (spectral counts) analysis were shown to be successful [3,4].

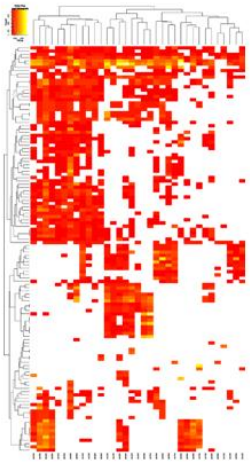


Figure 1. Original Data [3]

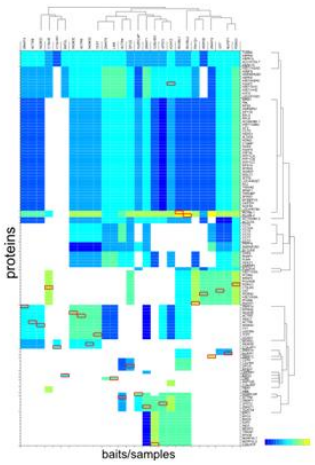


Figure 2. Biclustering

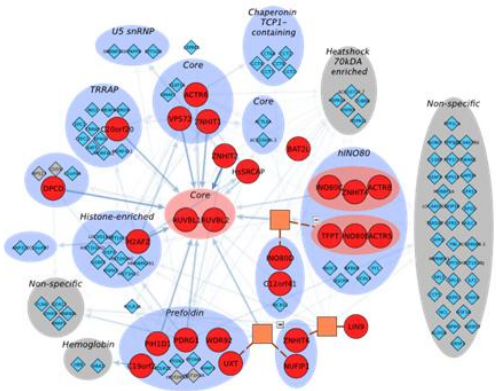


Figure 3. Network representation

The Bayesian checkerboard biclustering model for AP/MS data, that we had developed, combines both approaches and tries to decompose bipartite bait-prey interaction graph into non-overlapping bicliques (biclusters in matrix representation), while simultaneously requiring, that proteins within these biclusters have similar abundance levels. The resulting biclusters then correspond to interactions between protein modules/subcomplexes and co-complexed bait proteins.

The inference of the model is based on Markov Chain Monte Carlo (MCMC) sampling. To make sure that we effectively explore biclusterings space and obtain non-degenerated posterior distribution, advanced sampling techniques, such as equi-energy jumps, are used. The implementation could utilize the power of compute cluster, running as many MCMC instances in parallel as there are computational nodes available.

The method provides molecular biology expert with a distribution over all possible protein subcomplexes configurations, so that several hypotheses could be examined. The biological relevance of inferred biclusterings is additionally supported by the fact, that they improve the results of existing methods for filtering AP/MS datasets from contaminants and non-specific binders.

1. G.D. Bader, C.W. Hogue (2003) An automated method for finding molecular complexes in large protein interaction networks, *BMC Bioinformatics*, **4**:2
2. A.-C. Gavin et al. (2006) Proteome survey reveals the modularity of the yeast cell machinery, *Nature*, **440:**631–636.
3. M.Sardiu et al. (2008) Probabilistic assembly of human protein interaction networks from label-free quantitative proteomics, *PNAS*, **105:**1454–1459.
4. H.Choi (2010) Analysis of protein complexes through model-based biclustering of label-free quantitative AP-MS data, *Mol.Syst.Biol*, **6**:385.

The abundance of '2As' in different species.

Andriy Sukhodub, Lin Ruan, Gabrielė Stakaitytė, Garry Luke, Martin Ryan
University of St.Andrews, United Kingdom, as410@st-andrews.ac.uk

2A is an oligopeptide sequence first characterised in the picornavirus foot-and-mouth disease virus (FMDV). The FMDV genome comprises a single open reading frame encoding a polyprotein of ~ 2,300aa. The 2A region (just 18aa) delineates the upstream capsid protein domain of the polyprotein from the downstream replication proteins, in that it mediates a co-translational 'cleavage' at it's own C-terminus.

'2A-Like' sequences (2As) of viruses are found in the genomes of a range of mammalian, insect and crustacean positive sense and double-stranded RNA viruses and as was previously established in our lab they are all active. Whilst the –D(V/I)ExNPGP- motif is conserved amongst all 2As, the sequence upstream is not.

Our bioinformatic search has revealed ca 144 2As in genome of Stronglocentrotus purpuratus, 138 2As in the genome of Strongylocentrotus franciscanus and 291 2As in the genome of Allocentrotus fragilis. Also 2A sequences have been found in the genomes of Salmo salar (13), Ixodes scapularis (140), Acropora millepora (16), Biomphalaria grabrata (3), Caerorhabditis elegans (6), Saccoglossus kowalewski (64), Nematostella vectensis (8) and some other species.

Further analysis has revealed that the variations in the conservative domain show the stark predominance of the DVET-NPGP sequence that constitutes 16.5% of the whole pool of 2As with the largest contribution from sea urchins. Sea urchins largely contribute to the other most abundant conservative domains: DVEL-NPGP sequence constituting 12.6% of the whole pool, DVER-NPGP – 5.2%, , DVES-NPGP – 4%, DVEI-NPGP – 2%, DIHP-NPGP found in sea urchins only and DVHP-NPGP, found in sea urchins and Saccoglossus kowalewski – both 2.6%, DVEV-NPGP – 6.3% with the largest contribution from Sus Scrofa, , DIET-NPGP– 4.4% and DIEL-NPGP – 2.4% with the largest contribution from Ixodes Scapularis and other sequences less then 2%.

The evolutionary traces reveal 49 evolutionary 2A groups with different conservative domains and upstream context.

It is also found that most of the 2As, but not all of them display similar hydropathy patterns, which can explain their role in the suggested ribosome skipping mechanism.

To understand the functional role of found 2As and the variety of evolutionary traces we have constructed a number of vectors that contain viral TAV-2A known to have a strong cleavage function as well as fluorescent proteins (GFP and cherry) targeted to the different cellular parts. Cloning of representative 2As from the whole variety of the sequences found will allow us to answer the question about the role of 2As from different species in cellular signal pathways.

Detecting genes with triplet periodicity splicing

Yulia M. Suvorova, Eugene V. Korotkov

Centre of Bioengineering, Russian Academy of Sciences, 117312,
Russia, Moscow, Prospect 60-tya Oktyabrya, 7/1, suvorovay@gmail.com

The triplet periodicity (TP) is a common property of protein coding sequences and it is associated with a gene reading frame (RF). There are complex genes with more than one TP type along their sequence. We say that these genes contain a triplet periodicity splicing (TPS). The aim of our work is to find and study such genes.

We have developed a mathematical method to identify TP changes along a sequence; our measure is based on comparison of frequency matrixes of corresponding subsequences. Consider a coding sequence $S=\{s(k), k=1,2,...,L\}$, from the alphabet $A=\{a,t,c,g\}$. Let us introduce three RF in S and designate them as T_1, T_2 and T_3. RF T_1 really exists in the sequence while RFs T_2 and T_3 are the hypothetical ones.

Then let us introduce three matrices of the TP $M_1(x_1, x_2)$, $M_2(x_1, x_2)$ and $M_3(x_1, x_2)$, which are the matrices of TP calculated for RF T_1, T_2 and T_3 for the region of a sequence S from the base x_1 to the base x_2 ($S(x_1, x_2)$). An element m_{ij} of such a matrix shows the number of the base of a type i in position j of the codon (j=1,2,3) in the sequence. We should define the conditions which show us the existence of TPS after the x base in S. Firstly, TP should exist in the sequence S. Conditions of TP presence and quantitative measure for revealing TP are described in detail in work [8]. Secondly, we need a quantitative measure of a difference between two TP matrixes.

Let us assume that two TP matrices are significantly different if $U(M_1, M_2) \geq U_0$. Thus we define the condition of TPS presence in the position x as:

$$U\{M_1(1, x), M_i(x+1, L)\} > U_0 (i=1,2,3), \quad (1)$$

If a TPS occurs in x, then the matrix $M_1(x-L_1+1, x)$ is supposed to be different from $M_1(x+1, x+L_1)$, $M_2(x+2, x+L_1+1)$ and $M_3(x+3, x+L_1+2)$ matrixes.

For each TP matrix we calculated new one, which elements have normal distribution (matrix V_1 corresponds to $M_1(x-L_1+1, x)$, W_1, W_2 and W_3). The difference between the matrix V_1 and one of W_k matrix (k=1,2,3) is defined as: $U = D(1,k) = \sum_{i=1}^{4}\sum_{j=1}^{3}\left(\frac{v_1(i,j)-w_k(i,j)}{\sqrt{2}}\right)^2 (k=1,2,3)$

$D(1,k)$ is distributed as χ^2 with 6 degrees of freedom, if compared matrixes are calculated for random sequences. Following the conditions (1) we chose minimum of three value $D(1,k)$, k=1,2,3 ($D_{min.}$). The probability is $P(D_{min} \geq x) = P^3(\chi^2(6) \geq x)$. The final measure is defined as $F = -log_{10}P(D_{min} \geq x) = -3log_{10}(P(\chi^2(6) \geq x))$. Then we introduce a threshold value for F as $F_0 = -3log_{10}(P(\chi^2(6) \geq U_0))$.

We mowed a sliding pointer x along S. For each x we have varied the length of considering subsequences L_1 (from 60 to 600 nt), with a step of variation equal to 3 bases, we searched for L_1 that maximized measure F for the position. We select the point at which F reaches its maximum value and then checked its statistical significance. If this value is greater than threshold F_0, then one can consider that S has a TPS at the position x. Using our method we have analyzed every sequence from KEGG/Genes (release 48) databank [2]. The total number of genes with $F>F_0$ was 311221 (number of false positives was less than 5%).

We found (using [3]) that the repetitive or low-complexity sequences are not the only cause of TPS. Only about 7% of the whole result set are supposed to be caused by repetitive regions. We believe that the TPS cases may indicate a fusion of genes or domains in the region. We performed BLAST analysis to find potential ancestral genes for the parts of genes with TPS. As a result we found that in 131323 cases sequences with TPS have proper similarities (in different genes) for one or both parts (i.e. before and/or after the TPS point).

1. E.V.Korotkov, M.A.Korotkova. (2010) Study of the Triplet Periodicity Phase shifts in Genes *Journal of Integrative Bioinformatics*. 7:131-142.
2. M.Kanehisa, S.Goto. (2008) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*. 28:27-30.
3. L.Xuehui, T.Kahveci. (2006) A Novel Algorithm for Identifying Low-comlexity Regions in a Protein Sequence. *Bioinformatics*. 22:2980-2987.

Effect of HSV and L*a*b* Color Spaces on Segmenting Histological Images by Expectation Maximization Algorithm

Siamak Tafavogh

Center of Quantum Computation and Intelligent Systems, Faculty of Engineering and Information Technology, University of Technology Sydney., Australia, siamak.tafavogh@student.uts.edu.au

Neuroblastoma is an emberional tumor, derived from the neural crest and is tumors of sympathetic nervous system [1]. Classifying neuroblastoma based on shimada method [1] requires identifying the morphological features of Neuroblastic Tumors (NTs) such as neulei-cytoplasm, neuropil, mitosis karyorrhexis, and mitotic rate by pathologists. Image segmentation techniques allow computer to identify and extract the desire regions of the image. Digital slide scanners allow computerized analysis of neuroblastic tumor tissue specimens. The output of scanner is histological image in RGB color space. RGB is not perceptually uniform space. As a result we convert the RGB to uniform or approximately uniform color space such as L*a*b* and HSV color space respectively. In order to extract the morphological features such as nuclei-cytoplasm and neuropil of NT tissue specimen image, segmentation technique based on

Expectation Maximization (EM) clustering algorithm has been applied. Hematoxylin Eosin (H&E) stained specimens presents morphological features of the tissue by three colors: blue for nuclei-cytoplasm, pink for neuropil and white for background components. Therefore three clusters are considered for the EM algorithm. Segmenting histological images of NT tissue specimens are implemented by two different experiments. In the first experiment EM is applied to images in L*a*b* and HSV color space to group the identical histological components into the same cluster. In the second experiment EM is applied to the histological images of H&E stained tissue specimens, which are enhanced by Histogram Equalization (HE) techniques. 61 NT images are selected by a group of pathologists, and to evaluate the accuracy of segmentation Leave One Out method is implemented. 60 images are used as the training set and the excluded image is allocated to the test set. As a result a clustering model with three clusters is constructed by all existing pixels in the training set. Pixels of the test sets are clustered based on the clusters configuration derived from the model. The correctness of segmentation is compared against the "Oracle". The Oracle consists of 20000 nuclei-cytoplasm (blue), 20000 neuropil (pink) and 5000 background components pixels (white), were manually selected. In both experiments Precision and Recall Metrics evaluate the accuracy of segmentation. Friedman Aligned Ranked Test [24], which is an intra set comparison and Holm statistical hypothesis tests are used for comparing HSV and L*a*b* color spaces. Statistical tests based on the results of Precision and Recall in both experiments (applied HE and non applied HE to histological images) prove that EM in HSV color space achieves higher segmentation performance for nuclei-cytoplasm and neuropil in comparison with L*a*b*. Nuclei-cytoplasm and neuropil are the most significant morphological features for diagnosing neuroblastoma. Comparing performance of EM in both experiments clarifies that HE cannot improve the accuracy of EM algorithm in any of color spaces, while Holm and Friedman tests suggest that applying EM to non-enhanced histological images produces higher segmentation accuracy in both color spaces.

Our comparisons and statistics demonstrate that HSV color space boosts segmentation accuracy of EM for nuclei-cytoplasm and neuropil. Moreover, enhancing images by histogram equalization techniques cannot improve the accuracy of EM algorithm, while it has negative impact.

1. H. Shimada, I. Ambros, L. Dehner, J. Hata, V. Joshi, and B. Roald, "Terminology and morphologic criteria of neuroblastic tumors," Cancer, vol. 86, no. 2, pp. 349–363, 1999.
2. D. Montgomery and G. Runger, APPLIED STATISTICS AND PROBABILITY FOR ENGINEERS, With CD. Wiley-India, 2007.
3. J. Hodges and E. Lehmann, "Rank methods for combination of independent experiments in analysis of variance," The Annals of Mathematical Statistics, pp. 482–497, 1962.

Experimental Evidence of Optimal Interfacing of Subantennae in Superantenna of the Green Photosynthetic Bacterium *Oscillochloris trichoides* from the family *Oscillochloridaceae*

A.S. TAISOVA, E.P. LUKASHEV, N.V. FEDOROVA, A.V. ZOBOVA ,
L.A. BARATOVA, and Z.G. FETISOVA
M.V. Lomonosov Moscow State University, Moscow, 119992, Russian Federation
taisova@genebee.msu.ru; zfetisova@genebee.msu.su

This work continues the cycle of studies of the strategy of efficient functioning of natural light-harvesting antennae in accordance with our earlier concept on the strict optimization of the photosynthetic apparatus by functional criteria. Theoretical investigation of optimality of a model antenna functioning is a powerful tool for the study of efficient strategies for the light-harvesing in photosynthesis. Recently, on the basis of such model calculations, we postulated the biological expedience of existence in the *Osc.trichoides* superantenna of a new intermediate BChl *a* subantenna Bx that must ensure optimal interfacing of the chlorosomal BChl *c* subantenna B750 with the membrane BChl *a* subantenna B805–860 [1]. Targeted search for the theoretically predicted BChl *a* subantenna Bx allowed us to recognize it in *Osc.trichoides* chlorosomal baseplate subantenna.

It is well known that the baseplate subantennae in chlorosomes of green anoxygenic photosynthetic bacteria, belonging to both other families, *Chloroflexaceae* and *Chlorobiaceae*, represent a complex of BChl *a* with the ~6 kDa CsmA proteins. The idea of association of BChl *a* with protein in chlorosomes of *Osc. trichoides* was probed by low-temperature fluorescence spectroscopy and sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) analysis of alkaline-treated and native chlorosomes of *Osc. trichoides*. Alkaline treatment of *Osc. trichoides* chlorosomes resulted in disappearance of a BChl *a* band in their fluorescence spectra. Absorption bands of BChl *c*, the main light-harvesting pigment in *Osc. trichoides* chlorosomes, were not practically affected by alkaline treatment. In both *Osc.trichoides* and *Cf. aurantiacus* chlorosomes, only small changes in BChl *c* fluorescence were observed upon alkaline treatment at room temperature under both reducing and aerobic conditions. Determination of BChl *c* and BChl *a* content confirmed the removal of BChl *a* from *Osc. trichoides* chlorosomes upon alkaline treatment. We conclude that alkaline treatment of chlorosomes destroys BChl *a* in the baseplate while leaving BChl *c* in a form that is spectrally indistinguishable from that in untreated chlorosomes.

It was shown that upon alkaline treatment, only the 5.7 kDa CsmA protein was removed from the chlorosomes among five proteins detected by SDS-PAGE analysis, concomitantly with the disappearance of BChl *a* fluorescence emission at 821 nm (measured at 77K). Based on these results, we suggest that the baseplate BChl *a* subantenna does exist in *Oscillochloridaceae* chlorosomes as a complex of BChl *a* with the 5.7 kDa CsmA protein.

The present results support the idea that the baseplate subantenna, representing a complex of BChl *a* with a ~6 kDa CsmA protein, is a universal interface between the BChl *c* subantenna of chlorosomes and the light-harvesting BChl *a* subantenna of the cytoplasmic membrane in all chlorosome-containing photosynthetic bacteria.

The work was partially supported by the Russian Foundation for Basic Research.

1. A. V. Zobova, A. S. Taisova, and Z. G. Fetisova. Search for an Optimal Interfacing of Subantennae in Superantenna of Photosynthetic Green Bacteria from *Oscillochloridaceae* Family: Model Calculations. Doklady Biochemistry and Biophysics, 2010, Vol. 433, pp. 148–151.

Identification of Sources of Error Affecting Base Calling in Next Generation Illumina/Solexa Sequencing

Rene te Boekhorst[1], Irina Abnizova[2], Silvia Beka[1], Sandeep Brar[1], Imrana Sabir[1]

[1]*University of Hertfordshire, UK, r.teboekhorst@hets.ac.uk;* [2]*Wellcome Srust Sanger Institute, UK*

Motivation and Aim: In Illumina/Solexa sequencing clusters of single stranded cloned DNA fragments are attached to a flow cell (divided and sub-divided in "lanes" and "tiles"). The template DNA strands are complemented at each position by fluorescently labelled nucleotides in separate chemistry cycles. The fluorescence intensity of the clusters is measured for each of the four nucleotides and, ideally, at each cycle only one of the four signals is displayed. However, due to a number noise factors ("Phasing" and "Cross-talk", [1, 2]) this is mostly not the case and measures are needed to capture the (un-) ambiguousness of base-calling. These measures can be used in calibration, if they correlate well with the probability by which called bases are found back after aligning the read fragments (in which they are contained) to a reference genome. We investigated: 1) which measure best predicted the correctness of base-calling and 2) how much of the variance of this measure is due to lanes, tiles, cycles, identity of the called-base-call and its neighbours ("sequence context").

Methods and Algorithms: Eight metrics routinely used in ecology as measures of species diversity were adapted to assess the signal diversity of a base call. The metrics were computed for data obtained from the genome of the phage FX174, and sequenced by Illumina's Genome Analyzer GA2, release 1.4, run 3259. They were then correlated to the log odds of correct base calls by means of logistic regression. ANOVA was applied to the best performing measure to test for differences between tiles, lanes, type of nucleotide and cycle number. A second ANOVA was carried out to assess the effects of sequence context on the signal diversity of the middle nucleotide of a trimer.

Results: "Purity" (= maximum intensity/sum of the intensities) and a version of it (the relative sum of the two highest intensities) correlated best with the proportion of back-aligned base pairs. Lanes, cycles, and bases on explained up to 9% of the total variance of purity. The interaction of all four factors was significant and accounted for 11% of the overall variance. When T is the middle base of a trimer, preceded by G, average purity is lower than for all other combinations.

Conclusion: Illumina's original metric (purity) is a good measure for sequencing accuracy, but its value depends significantly on the combined effect of lane, tile, cycle and nucleotide.

1. J.C. Dohms et al. (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing, Nucleic Acids Research, 36: 1-10.
Y. Ehrlich et al. (2008) Alta-cyclic: a self-optimizing base caller for next-generation sequencing. Nat Methods, 8: 679-82.

Heterotachy of double substitutions in neighboring nucleotides in non-coding sequence

Nadezhda Terekhanova[1], Alexey Kondrashov[2], Georgii Bazykin[1]

[1]*Institute for Information Transmission Problems of the Russian Academy of Sciences, Russian Federation,* *nadterekhanova@yandex.ru*
[2]*University of Michigan*

Both the mutation rate and the selective pressure are non-uniform along the non-coding DNA segments. However, the durability of this variation is poorly understood. Here, we study the tendency of pairs of neighboring substitutions in the non-coding sequence to occur in the same phylogenetic lineage. The lineage in which each substitution has occurred is revealed by maximum parsimony in comparison of human with chimp (using orangutan as an outgroup), and in comparison of D. melanogaster and D. simulans (using D. erecta as an outgroup). In both cases, the substitutions in neighboring nucleotides tend to occur in the same lineage. This effect decreases with distance between nucleotides, but is still pronounced for substitutions at ~10 nucleotides from each other. Possible mutational and selective explanations for this pattern are discussed.

Properties of Intronic miRNAs Affecting CDH1 Gene

Anatoliy T. IVACHSHENKO, Olga A. BERILLO, Vladimir A. KHAILENKO

Kazakh National University named Al-Farabi, Kazakhstan, a_ivashchenko@mail.ru

Intronic miRNAs represent significant interest for revealing post-transcriptional regulation of protein-coding genes expression. These miRNAs can come directly from DNA transcripts of introns of protein-coding genes (pre-miRNA) in the presence of promotors in the miRNA gene. Another way of maturing miRNA is cutting from intron of pre-miRNA before or after splicing. Intronic miRNAs can affect on mRNA of protein-coding genes and on mRNA of host gene. Possible miRNA effect on mRNA of host gene represents the important example of self-control of gene expression. Participation of protein coded by such genes by means of intronic miRNA will influence metabolic ways including these enzymes. Studying of properties of intronic miRNAs in comparison with intergene miRNAs presents great interest. Intronic miRNAs effect on mRNA of host gene and CDH1 gene participating in development of cancer of esophagus, stomach, intestines and breast of human have been studied.

miRNAs acting on mRNA of CDH1 gene were borrowed from a database miRBase. The energy interaction (ΔG) of 15 intronic miRNAs with mRNA of CDH1 gene was calculated. The various degrees of their linkage were showed. For each intronic miRNA from two to four sites of linkage with mRNA of CDH1 gene were revealed. The number of sites of interaction of miRNA with mRNA was defined on the basis of value ΔG and its standard deviation for each site with level reliability $p<0.001$ was calculated. The revealed sites of action miRNA were characterized by ΔG value which was above 50% of the maximum ΔG value possible for miRNA. In miRBase intronic miR-590-3p had no sites of linkage with mRNA of CDH1 gene. However average ΔG value of first five supposed sites of linkage made up -16.7±3.0 kcal/mol and for certain had no difference from average value ΔG (-16.0±2.1 kcal/mol) for interaction mRNA-mRNA in 2D-structure.

Twelve intronic miRNAs interacted with mRNA of host gene significantly. The ΔG values for these miRNAs had above 50% increase of the maximum ΔG value possible for particular miRNA. The number of sites of interaction changed from one (miR-98, let-7g, miR-185) to five (miR-224). These data indicate the ability of intronic miRNAs to compete with intergenic miRNAs in regulation of genes expression. mRNA of AAKT, EIF4H and RNF130 genes, coding miR-338-3p, miR-590-3p and miR-340 accordingly, had no sites of binding for the intronic miRNAs. Therefore, these genes have no self-control of expression by means of their own miRNAs.

Intronic miRNAs of host gene contact with 5'UTR, CDS and 3'UTR of its mRNA. So, among 29 sites of binding of intronic miRNAs with mRNA of host genes in 5'UTR there were four sites, in CDS - ten sites and in 3'UTR - fifteen sites. In mRNA FGF13 gene all sites for miR-504 were disposed in 5'UTR. In mRNA of EML2, PDE2A and HUWE1 genes (for miR-98) all sites for corresponding miRNA are in CDS.

Among 39 sites of binding of intronic miRNAs with mRNA of CDH1 gene in CDS there were 19 sites and in 3'UTR - 20 sites. In 5'UTR of mRNA of CDH1 gene there were no sites of linkage for all intronic miRNAs. Action sites miR-23b were disposed only in coding sequence mRNA of CDH1 gene. Sites of miR-98 and miR-340 were disposed only in 3'UTR.

Existence of equal quantities of binding sites miRNA in CDS and 3'UTR were showed. These results indicate suppression of expression of target genes. Thus it is necessary to consider not only their interaction in 3'UTR. For example, of let-7g, miR-139-5p, miR-185, miR-338-3p, miR-504 energy of interaction in sites CDS of CDH1 gene was higher, than for sites located in 3'UTR mRNA.

Feature of Interaction of Intergenic miRNAs with mRNA of CDH1 Gene Participating in Development of Cancer

Asel S. ISSABEKOVA[1], Anatoliy T. IVACHSHENKO[1], Mireille REGNIER[2]

[1]*Kazakh National University named Al-Farabi, Kazakhstan, a_ivashchenko@mail.ru*
[2]*INRIA, Paris, France, mireille.regnier@inria.fr*

With the establishment of the role of miRNA in gene expression regulation at post-transcription stage, a significant success in detection biological value of this molecule and mechanisms of their influence was achieved. Human miRNA databases consist of more than 1000 validated genes and every miRNA effects on several protein-coding genes. That evidence confirms their direct or mediated regulation of expression of a significant part of human genome. Mutations in miRNA genes or in binding sites in mRNA lead to different pathology. A miRNA participation in the development of different types of cancer has been shown. For many miRNAs, correlative changes have been established between the miRNA level and the cancer localization, their morphological types, the expression of oncogenes and tumor-suppressor genes. An active research is ongoing on the feature of miRNA effect on the expression of protein-coding genes that participate in oncogenesis.

This work studies the feature of interaction of intergenic miRNA with mRNA of CDH1, a gene that participates in the development stomach, esophageal, colon and breast cancer. Nucleotide sequences of 25 miRNAs were obtained from miRBase databases. For each miRNA, a maximal energy of binding may be defined and computed with (modified) program RNAhybryd. Given a miRNA, for each position where miRNA and mRNA may bind, RNAhybrid computes a binding energy (ΔG). This value ΔG may change significantly and was up to 68% from maximal energy of binding. A feature of animal miRNA is the similarity of binding energy with the best mRNA sites. A possible interpretation is that it allows mRNA translation to be completely inhibited when miRNA is abundant. A contrary, plants have one leading binding-site of miRNA with mRNA. Therefore, we only analyze the best binding sites, one to four, as long as the differences between their ΔG are not significant. A widespread opinion is that miRNA bind with mRNA only in 3'UTR. This study shows that exist potential binding sites with high energy in 5'UTR, CDS and 3'UTR mRNA of CDH1 gene. Moreover, the distribution of binding sites length does not depend of the position in mRNA.

For instance, for miR-23a, the three best binding sites are found in CDS of mRNA CDH1 gene. For miR-485-5p, the best binding site is in 5'UTR and the two following sites are in 3'UTR. For miR-296-3p, the three best binding sites are in 5'UTR, CDS and 3'UTR.

Meanwhile, the value of ΔG for different miRNAs changed from -20.3 to -32.2 kcal/mol. Interaction energy of mRNA-mRNA in 2D structure of mRNA of CDH1 gene was computed on the whole mRNA. A division by miRNA length provides normalization, leading to an average value -12.96; -13.6; -14.3; -14.9 kcal/mol when the miRNA length is 20-23 nt. Differences were computed for each miRNA and extreme values are found for length 22 nt, ranging from -6 to -17.9 kcal/mol.

Conclusion: Obtained results show several binding sites for one miRNA. Binding sites of intergenic miRNA with mRNA of CDH1 gene localize in 5'UTR, CDS and 3'UTR.

Peculiarities of Interaction mir156a and mir396a Arabidopsis thaliana with mRNA their Target Genes

Asyl A. BARI, Shara A. ATAMBAYEVA, Anatoliy T. IVACHSHENKO
Kazakh National University named Al-Farabi, Kazakhstan, a_ivashchenko@mail.ru

Plant microRNAs (miRNA) play key roles in regulating many major biological processes. Their expression in stem, root, leaves and flowers is revealed. Participation miRNAs in development of plants is shown. Change of their concentration is noted at the reply of plants to various stressful factors: heavy metals, temperature, drought, salinity etc. Studying of properties and a role miRNA in these processes represents unconditional interest as they can change considerably a current of key biological processes and influence stability and efficiency of plants. Despite increasing number of publications on plant miRNAs, many questions of their interaction with mRNA target genes remain not studied. Now in A. thaliana it is identified more than 200 miRNAs and for each of them interaction with mRNA target genes are established. It is obviously important to elicit what sites mRNA (5'UTR, CDS, 3'UTR) operate miRNA and degree of binding miRNA with these sites. Presence in mRNA one gene of several sites of interaction with miRNA as it is observed in animal objects is possible. For the majority plant miRNA these and other questions remain insufficiently studied.

Expression level miR156a and miR396a in A. thaliana strongly changes in reply to salt and oxidizing stress, accordingly. In this connection we have studied degree of linkage of these miRNA with mRNA target genes A. thaliana resulted in miRBase. For all target genes one site of interaction miRNA with mRNA was distinctly allocated. Energy of interaction (ΔG) in other sites was up to energy of interaction mRNA-mRNA and, hence, it is possible to consider, that for everyone miRNA there is only one site of interaction that testifies to sufficiency of one site for blocking of translation of mRNA-target. Energy of interaction between miR156a and eleven

mRNAs its target genes changed from -35.2 to -38.9 kcal/mol (it was on the average equaled -37.5±1.20 kcal/mol). These changes ΔG made from 84 to 93% from maximum energy equal -41.8 kcal/mol, linkage miRNA with completely complementary to it RNA. Average energy of interaction mRNA-mRNA at formation 2D structures of genes counting on 22 nt. (the length miR156a) varied from -16 to -20 kcal/mol, that is there essentially less ΔG values of interaction miRNA-mRNA. To consider the contribution of randomly interaction miRNA with mRNA we have calculated energy of interaction miR396a with miRNA target genes for miR156a which on the average it was equaled -22.4±2.0 kcal/mol. Average energy of interaction miR156a with mRNA target genes for miR396a was equaled -22.4±2.4 kcal/mol.

Energy of interaction miR156a with mRNA AT1G53160 gene was equaled -20.9 kcal/mol. Proceeding from this value and ΔG values casual interaction miRNAs with mRNAs we have drawn a conclusion, that mRNA AT1G53160 is not a miR156a target. Energy of interaction miR156a with mRNA AT2G21840 made -30.7 kcal/mol, that testifies to presence in this mRNA concerning a weak site of linkage.

Energy of interaction between miR396a and thirteen mRNAs its target genes in the first site on the average made -31.9±1.2 kcal/mol. Energy of interaction of the second site was essentially more low and was equaled -21.5±1.2 kcal/mol. This value is comparable with energy of randomly interaction (-22.4±2.4 or -22.4±2.0 kcal/mol) and hence in mRNA all studied target genes sites in the second position are not competitive to sites in the first position.

Interaction sites miR156a with mRNAs are localized mainly in CDS and only in mRNA AT3G15270, AT1G27370 and AT2G33810 genes in 3'UTR and mRNA AT3G28690 gene in 5'UTR. Interaction sites on coding area mRNA target genes for miR396a are randomly distributed, that is, there is no their preferable arrangement by 3'-end or by 5'-end CDS. In mRNA AT3G15270 and AT2G33810 sites are located in 3'UTR, and in mRNA AT3G28690 gene in 5'UTR. In the range of 30-90% of length CDS mRNA all genes 95% of sites without a maximum in distribution are located.

In miR156a at position 4 from 5'-ends there is a triplet 5'CAG which in ribozymes promotes splitting complement RNA on a triplet 5'CUG. In miR396a this triplet settles down from 6th position. Frequency of occurrence 5'CAG in 217 miRNAs A. thaliana it was equaled 0.272 and was essentially above expected equal 0.115. Hence, this triplet has the raised frequency not casually. From 217 miRNAs A. thaliana 59 miRNAs had 5'CAG and it the triplet met in 4, 7 and 9 positions miRNAs is more often.

Approach for Gene Networks Phylogenetic Decomposition

Vladimir Timonov, Konstantin Gunbin, Igor Turnaev
Institute of Cytology and Genetics SB RAS, Russian Federation, vtimonov@bionet.nsc.ru

There are huge amount of gene superfamilies sufficiently well studied [1]. However, the particular gene function only makes sense in the context of gene network controlling certain cellular function. The world literature has accumulated huge amounts of data on various gene expression properties and on various associations between mutant alleles and phenotypical abnormalities. This provides great opportunities for semiautomatic analysis of gene network molecular evolution.

The server-side pipeline and the graphic tool for analyzing the molecular evolution of gene networks were developed. Our approach is based on the protein-coding gene duplication analysis [2], so we named our computer system as "Phylogenetic Decomposition (PD)". PD is the sequential joining of gene network node pairs (paralogous genes/proteins). Joining order is determined in accordance with the duplication number from the top of the phylogenetic tree of certain protein superfamily to its root. The results of nodes joining can be visualized and manipulated in a very flexible graphical subsystem.

Our system is based on client-server architecture. The phylogenetic decomposition pipeline on the server-side provides all phylogenetic computations. The client-side software tool was developed and named GeneNetStudio. The GeneNetStudio application provides gene network visualization, manipulation and analysis. To implement PD pipeline we have created our own simple client-server platform. The first automatic pipeline chain provides collecting of gene network proteins and its clustering: proteins collected through ID-based searches against GenBank NR database, and then clustered by BLASTCLUST tool. The second automatic pipeline chain processes five phylogenetic analysis steps. On the final fifth step of this chain user can define the duplication events on which paralogous protein functions could dramatically changed.

The gene network of a mammal's cellular cycle was taken for system testing [3]. It is off known that the regulation core of this gene network consists of 5 protein superfamilies, such as Cyclins, CDK, E2F, pRB and CDI [4]. We carried out phylogenetic decomposition for animal cell cycle gene network consisting of these protein superfamilies. We received only 22 directed regulatory circuits after carrying out greedy phylogenetic decomposition. The stepwise phylogenetic decomposition of the gene network of animal cell cycle resulted that over 80% of evolutionary events in this gene network appears to be a kind of gene duplications followed by origin of competitive inhibitor/activator protein of ancestral function. This finding describes one

more important way for increasing the adaptive plasticity of the organism through duplications of multi-functional proteins [5].

Thus, the specialized method and software package consists of server and specialized client visual module has been developed. The tools intended to build a hypothetical ancestral gene networks. Such networks can be useful to the functional analysis of extant gene networks due to its big size and complexity. Software prototypes are available on http://pixie.bionet.nsc.ru/samem/phylodecompose.html.

Research was supported by the RFBR grant №09-04-01641-a; integration projects of the Siberian Branch of RAS № 113, 119; programs of RAS №6.8, №B26.29, №24.2; contract of the Russian Ministry of education and science №857.

1. KEGG Pathway DB, http://www.genome.jp/kegg/pathway.html
2. S.A. Teichmann, M.M. Babu (2004) Gene regulatory network growth by duplication, Nat. Genet. 36: 492-496.
3. I.I. Turnaev, O.A. Podkolodnaya (2002) Gene network of a cell cycle control. Proceedings of the 3rd international conference on Bioinformatics of Genome Regulation and Structure, 2: 95-98.
4. P. Kaldis (2006) Cell Cycle Regulation. (Springer-Verlag).
5. W. Ma, A. Trusina, H. El-Samad, W.A. Lim, C. Tang (2009) Defining network topologies that can achieve biochemical adaptation. Cell. 138: 760-773.

Construction and expression in E. coli and cyanobacteria of the deletion derivatives of the cyanobacterium Synechocystis sp. PCC 6803 drgA gene and its hybrids with gfp

Victoria A. TOPOROVA[1], Alexandr V. ALESHIN[2], Alexey N. NEKRASOV[1], Elena M. MURONETS[2], E.P. LUKASHEV[2], K.N. TIMOFEEV[2], Dmitry A. DOLGIKH[1] AND Irina V. ELANSKAYA[2]

[1] *Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Russian Academy of Sciences, Moscow, Russia*

[2] *Department of Genetics, Faculty of Biology, Lomonosov Moscow State University, Moscow*

toporova-viktorija@rambler.ru

Soluble NAD(P)H:quinone-oxidoreductase encoded by *drgA* gene of the cyanobacterium *Synechocystis* sp. PCC 6803 is involved in NADPH oxidation and is responsible for the cell sensitivity to nitroaromatic inhibitors as well as for the resistance to the oxidative stress inducer menadione [1]. DrgA protein is responsible for peroxide reduction in Fenton reaction [2] and participates in regulation of photosynthetic and respiratory electron transport in cyanobacterial thylakoid membranes [3].

The protein sequences of DrgA from *Synechocystis* sp. PCC 6803 and its homologues from other microorganisms were aligned and studied for their information content by analysis of Shannon-Weaver informational entropy computed as function of the distance between the amino acid residues [4-6]. Sites of increased degree of information coordination between residues (IDIC-sites) were identified. Associations of information-coordinated structural elements (IDIC-trees and IDIC-branches) were mapped.

Coding sequence of *drgA* gene was amplified using PCR method. To study DrgA functional topology, several new deletion derivatives of *drgA* gene (*drgA∂1*, *drgA∂2*, *drgA∂3*, *drgA∂4* and *drgA∂5*) were constructed using PCR. In order to facilitate protein purification we have spliced the 3'-ends of all genes with 12xHis tag coding sequence. For visualization of DrgA, the genes encoding the green fluorescent proteins (GFP) *cherry* or *egfp* were placed between *drgA* and 12xHis tag coding sequences. Several constructions for direct constitutive and inducible intracellular expression in *E. coli* of *drgA* and its deletion variants were designed and investigated. The recombinant proteins were purified by IMAC-chromatography method. The enzyme activity of DrgA was tested. The purified DrgA-12His protein exhibited high quinone reductase and nitroreductase activity. The rate of re-reduction of photooxidized Photosystem I reaction center was increased after addition of DrgA-12His protein and NADPH to isolated cyanobacterial thylakoid membranes. Thus, DrgA protein may participate in electron transfer from NADPH to plastoquinone pool in thylakoid membranes of the cyanobacterium *Synechocystis* sp. PCC 6803.

To study the role and cellular localization of DrgA in Synechocystis sp. PCC 6803 cells on the basis of the constructed *drgA* deletion variants the clones of Synechocystis sp. PCC 6803, carrying deletions in selected regions of a *drgA* gene, as well as clones of cyanobacteria that contain protein DrgA hybrids with Cherry and EGFP were obtained and tested for their photosynthetic activity and resistance to inhibitors. Also the Synechocystis sp. PCC 6803 cells with spontaneous mutations in *drgA* gene were obtained and characterized.

The work was supported by RFBR grant 09-04-01119.

1. Elanskaya I.V., Chesnavichene E.A., Vernotte C., and Astier C. (1998) Resistance to nitrophenolic herbicides and metronidazole in the cyanobacterium *Synechocystis sp*. PCC 6803 as a result of the inactivation of a nitroreductase-like protein encoded by *drgA* gene. FEBS Letters, **428:** 188-192.
2. Takeda, K., Iizuka, M., Watanabe T., Nakagawa, J., Kawasaki, S., and Niimura Y. (2007) *Synechocystis* DrgA protein functioning as nitroreductase and ferric reductase is capable of catalyzing the Fenton reaction. FEBS J., **274**: 1318-1327.
3. Matsuo M., Endo T., and Asada K. (1998) Isolation of a novel NAD(P)H-quinone oxidoreductase from the cyanobacterium *Synechocystis* PCC 6803. Plant Cell Physiol., **39**: 751-755.
4. Nekrasov A.N. (2002) Entropy of Protein Sequences: an Integral Approach. Journal of Biomolecular Structure & Dynamics, **20**: 87-92.
5. Rogov S.I., Nekrasov A.N. (2001) A Numerical Measure of Amino Acid Residues Similarity Based on the Analysis of their Surroundings in Natural Protein Sequences. Protein Engineering, **14:** 459-463.

Pharmacogenetics of disease modifying treatment in Russian patients with multiple sclerosis

Olga G. Kulakova[1], Ekaterina Yu. Tsareva[2], Vitalina. V. Bashinskaya[1], Alexey N. Boyko[2], Sergey G. Shchur[2], Dmitry V. L'vov[3], Alexander V. Favorov[3], Olga O. Favorova[1]

[1]N.I. Pirogov Russian State Medical University, Russian Federation, kateritsa@gmail.com
[2]Moscow City Multiple Sclerosis Center, Russian Federation
3 Research Institute for Genetics and Selection of Industrial Microorganisms, RU, Russian Federation
4 Oncology Biostatistics and Bioinformatics, Johns Hopkins School of Medicine, Baltimore, US, United States

BACKGROUND. Multiple sclerosis (MS) is an inflammatory, demyelinating disorder of the central nervous system which is thought to be immune-mediated and exhibit varying response to disease modifying treatment (DMT). Immunomodulatory strategies are being actively involved in therapeutic intervention in MS. Long-term medication with specific first-line DMTs, glatiramer acetate (GA, Copaxone) or interferon-beta (IFNb), have been shown to reduce the number of relapses, to delay the new lesions formation and disability progression during MS. Despite DMT efficiency, which has been shown in numerous clinical trials, significant proportion of patients appears to have little benefit from GA or IFNb treatment. For MS patients who are resistant to one or another DMTs the opportunity to get an alternate treatment as early as possible is extremely important. Genetic variants, affecting on individual drug efficiency, could serve as biomarkers, which determine a choice of certain DMT for particular patient. In this study we analyzed the association of efficiency of response to GA and IFNb treatment in MS patients with the immune response genes' polymorphisms, which mainly code cytokine network components, in order to find a possibility of applying the genetic status of MS patients for selecting DMT in an early phase of the disease.

MATERIALS AND METHODS. DNA samples were obtained from 285 MS patients treated with GA and from 253 MS patients treated with IFNb. The response to DMT was considered to be optimal if there were no relapses and/or no sustained progression of EDSS level during the whole treatment period (no less than 2 years). Non-responders were defined as patients who have one or more relapses per year, required treatment with corticosteroids, and/or progression of the EDSS level (>= 1 point) in this period. The combinations of alleles of different loci (allelic combinations), which carriage was associated with response to DMT were identified using APSampler algorithm. All MS patients were genotyped at polymorphic loci of the following candidate genes: tumor necrosis factor (TNF, −308G>A, rs1800629), interferon-gamma (IFNG, 874T>A, rs2430561), transforming growth factor beta1 (TGFB1, −509C>T, rs1800469), interferon-beta (IFNB1, 153T>C, rs1051922), first subunit of CC-chemokine receptor-5 (CCR5 w→del32), interferon-alpha/beta receptor (IFNAR1 16725G>C, rs1012335),

alpha subunit of interleukin 7 receptor (IL7RA exon 6 C>T (Thr244Ile), rs6897932), cytotoxic T-lymphocyte antigen 4 (CTLA4, 49A>G, rs231775) and beta chain of HLA class II gene (DRB1) in responders (Rs) and non-responders (NRs).

RESULTS. We compared Rs versus NRs. Reliable associations of carriage of several allelic combinations of CCR5, DRB1, TGFB1 and IFNAR1 genes with poor response to GA treatment (OR<1) were observed. The most significant p-value refer to the combination including alleles of all above genes: DRB1*15+CCR5*d+TGFB1*T+IFNAR1*G (p=0.00018, OR=0.072). This association was found for men and women separately.

Some combinations including alleles of CCR5, IFNAR1, DRB1, TGFB1 and IFNG genes were associated with poor response to IFNb treatment. The most significant p-values refer to the combinations DRB1*16+CCR5*w/w (p=0.007, OR=0.1) and TGFB1*T+ IFNAR1*C+IFNG*A (p= 0.01, OR=0.5). Notably, some allelic combinations, associated with poor response to IFNb, included alleles CCR5*w, IFNAR1*C and DRB1*16, whereas allelic combinations, negatively associated with response to GA include alleles CCR5*d, IFNAR1*G and DRB1*15.

CONCLUSIONS. The results of present study may be useful for selecting the alternate DMT for the individual patient in an early phase of the disease according to his/her CCR5, IFNAR1 and DRB1 genotypes. This work was supported by RFBR and Marie Curie (UEPHA*MS) grants.

Phosphorylation Dynamics During the Cell Cycle Shows Preferences for Different Protein Structural Propensities

Stefka Tyanova[1], Juergen Cox[2], Dmitrij Frishman[1]

[1] *Genome-oriented Bioinformatics, Technische Universitaet Muenchen, 85354 Freising,Germany*
[2] *Proteomics and Signal Transduction, Max-Planck-Institute for Biochemistry, 82152 Martinsried,Germany*
tyanova@wzw.tum.de, *d.frishman@wzw.tum.de*

Rapidly evolving mass spectrometry-based technologies together with stable-isotope labelling techniques and advanced bioinformatics analysis provide powerful means for overcoming the problem of low abundance of phosphorylation events and result in high-resolution site-specific quantitative phosphorylation data [1].

In addition to the frequently applied linear motif-centric view, various structural properties of phosphorylation sites are reported to enhance recognition and specificity. The significance of intrinsic disorder has been demonstrated [2], however, not excluding cases in which defined structural regions turn out to be relevant.

Despite its importance for the regulation of a myriad of cellular process and the large number of studies focused on protein phosphorylation, no systematic research, which makes use of large scale quantitative phosphoproteomics data in combination with structural features, has been conducted. In this work, we try to bridge this gap and investigate the relation between changes in phosphorylation levels during 6 time points of the cell cycle, as measures by Olsen et al. [3] and structural propensities of the modified sites.

Our analysis suggests that sites, which are found within distinct structural environments, exhibit different phosphorylation dynamics. On one hand, modified residues, which lie within intrinsically disordered regions, are characterized by large variation in their occupancy during the cell cycle. High dynamics is synonymous to tight regulation and hence to functional importance of those modifications for the cell cycle progression, which underlines the significance of disorder regions. The association between lack of regular structure and highly variable phosphorylation is further supported by strong preferences for proline-directed kinases [4] as regulatory enzymes of sites with dynamical occupancy, revealing the presence of a regular structure breaker residue.

On the other hand, sites, which retain constant level of phosphorylation during the 6 time points, are predominantly found within regular secondary structures (alpha helices and beta sheets). Interestingly, variation in dynamics is inversely proportional to the average level of phosphorylation of the site and examples of constitutive phosphorylation are present. Unlike in the previous group, proline-directed kinases are underrepresented and instead basophilic and acidophilic kinases are preferred. This enrichment of charged residues in the surrounding of the modified residues could either facilitate the addition of negative charges (by stabilizing interactions with positively charged residues) or lead to structure disruption (by repulsive interactions with negatively charged residues).

In summary modifications of residues in regular structures are characterized by slower dynamics and long-lasting effect, while multiple phospho-sites in disorder regions have largely varying occupancy levels and have been hypothesised to have a gradient effect on the protein function. Therefore it would be worth studying how these properties influence rapid and robust cellular response to a given stimulus and what is their effect on free energy governing conformational changes.

References:

1. Cox J, Mann M (2007) *Cell* **130**.
2. Iakoucheva LM et al. (2004) *Nucleic acids research* **32**: 1037-1049.
3. Olsen JV et al. (2010) *Science signaling* **3**: ra3.
4. Zhu G et al. (2005) *The Journal of biological chemistry* **280**: 10743-10748.

GPGPU-assisted prediction of ion binding sites in proteins

Leonid Uroshlev[1], Sergei Rahmanov[1], Ivan Kulakovskiy[2], Vsevolod Makeev[1]

[1]*IOGen, Moscow, Russian Federation, leoniduroshlev@gmail.com*
[2]*Engelhardt Institute of Molecular Biology, Moscow, Russian Federation*

ABSTRACT. Prediction of binding sites for different types of ions in protein 3D structure context is a complex challenge for biophysical computational methods. One possible approach involves using empirical, also called as knowledge-based, potentials. In the current study, we present a new GPGPU program complex, PIONCA (Protein-ION CAlculator) for efficient generation of empirical potentials for protein-ion interaction, provide description of its characteristics, and also present a publically available online service based on it.

1. INTRODUCTION. Protein molecules are intensively studied due to their paramount role in the cell. Approximately 30% of all proteins in PDB are metalloproteins; ions act as cofactors in most redox reactions, participate in signaling, protein and nucleic acid folding, interaction, dimerisation, etc. Due to high cost and often inavailability of experimentation with proteins, modeling of the processes of protein interaction with water and intracellular ions is crucial to understand normal function such as catalysis, and in some cases, in pathology and treatment. Molecular dynamics routinely ignores angular aspects of ion coordination by protein atoms and is not always adequately representing the interaction, and also computationally costly. Here, we present a GPGPU implementation of PIONCA web service dedicated to this task.

2. METHODS. Our method is based on statistical potentials for atom interactions, derived from PDB using stochastic reference state as described in [1], [2]. Direct linear implementation of the algorithm, though, is rather time-consuming, due to a necessity to create and process large numbers (hundreds of thousands) of nodes in a 3D mesh encompassing the protein structure. Calculating the mesh on a regular workstation PC may take hours for medium to large proteins at high mesh resolutions. An obvious advantage exists, in connection with the fact that every node can be processed independently and in parallel with the others. Thus, it seemed optimal to use GPGPU-programming.

3. RESULTS. Nvidia CUDA was selected as GPGPU platform for parallel algorithm implementation, as the most popular and cost-efficient. We report ten times faster calculations using the GPGPU. Figure 1 shows the dependence of average calculation time versus structure size (number of amino acids).

1. Rakhmanov S.V., Makeev V.J. "Atomic hydration potentials using a Monte Carlo Reference State (MCRS) for protein solvation modeling". BMC Structural Biology 2007, 7:19
2. Rahmanov, Kulakovskiy I, Uroshlev L, Makeev V. (2010) Empirical potentials for ion binding in proteins. J Bioinform Comput Biol. Jun;8(3):427-35.

Proteogenomic annotation of recently sequenced bacteria strains

Alexey Uvarovskii[1], Dmitry Alexeev[2]

[1]*Moscow Institute of Physics and Technology, Russian Federation, alexey.mipt@gmail.com*
[2]*Institute of Physical Chemical Medicine, Russian Federation*

After developing of modern technologies DNA sequencing became routine procedure not only for microorganism, but for multicellular organisms, and genomic data amount increases exponentially. Along with this, efforts in proteomics of the last 10 years allow to gain complete bacterial proteomes, herewith the creation of those proteomic maps for the organism with few thousands of annotated genes takes about a week. An exception is only the proteins having extreme physical chemical properties (light molecular weight, large amount of transmembrane domains or nonstandard pI), but the rate of them is not bigger than 10% of the observing amount.

However, methods of the large data generation rise much more than technology of data processing. So, annotation even of the simplest microorganisms is complicated problem, which is so far not fully automated. Although automatic annotation based on computational techniques greatly simplify the solution, during the sequencing and ORF prediction errors may occur, and the final annotation may vary from one computational approach to other. These differences in genome data interpretation can affect protein identification results.

One of the modern approaches which allow increasing the annotation precision is proteogenomics. This approach consists in correction of annotation on the basis of proteomic data. By means of this technique one can prove presence of gene product, clarify N-chains and detect splice variants.

In this work we used proteomic data for genome annotation for Spiroplasma melliferum by virtue of sequence data for this organism and genome annotation for Spiroplasma citri. Also we produced similar work for Helicobacter pylori (strain a45) and compared its proteome and genome with Helicobacter pylori (strains J99 and 26695).

Functional conservation beyond sequence conservation

Olga Vakhrusheva[1,2], Georgii Bazykin[1,2], Alexey Kondrashov[1,3]

[1]*Department of Bioengineering and Bioinformatics, Moscow State University, , Moscow, Russia;*
[2]*Institute for Information Transmission Problems (Kharkevich Institute), Russian Academy of Sciences, , Moscow, Russia, vakh57@gmail.com*
[3]*Life Sciences Institute and Department of Ecology and Evolutionary Biology, University of Michigan, Michigan, USA, kondrash@umich.edu*

At moderate evolutionary distances, functional significance of noncoding sequences can be readily assessed through their above-random conservation between genomes. However, there is accumulating evidence that regulatory sequences are capable of rapid turnover, and, generally, conservation of sequence is not a necessary prerequisite for conservation of function.

For example, sequences at orthologous loci conserved among teleosts and mammals, but lacking significant sequence similarity between mammals and teleosts, can drive similar patterns of gene expression in zebrafish transgenic assays[1]. These findings may point to insufficiency of sequence similarity-based approaches to analysis of functional conservation. We used a bioinformatic approach to the question of whether functional conservation is possible without sequence conservation, and whether it can be inferred from genome wide comparative studies.

We considered two pairs of species chosen so that the evolutionary distance between species in a pair is significantly less than between species from two different pairs, though large enough to ensure that conservation in noncoding regions is mainly due to their functional role. We hypothesized that functional conservation between pairs of species should lead to above-average occurrence of conserved (within each pair) sequences at loci which are orthologous between pairs, even when no conservation between pairs is observed. Orthologous introns were chosen as sequence units within which conservation was analyzed, as they are short, naturally bounded by exons, and their orthology can be inferred even at large phylogenetic distances.

We selected introns orthologous in all the four species (defined as the introns in orthologous positions of coding sequences in orthologous proteins). The sequence similarity between orthololologous introns from different pairs was negligible. Nevertheless, there was a 2-4-fold excess of 4-sets of orthologous introns in which each species pair carried a conserved non-coding element, compared to the random expectation. These results imply that functional conservation of non-coding sequences is manifested at high evolutionary distances through presence of conserved segments of DNA at orthologous loci.

1. Fisher, S., Grice, E.A., Vinton, R.M., Bessling, S.L. & McCallion, A.S. Conservation of RET regulatory function from human to zebrafish without sequence similarity. *Science* **312**, 276-279 (2006).

Flexible and robust patterning by centralized gene networks

Sergey Vakulenko[1], Ovidiu Radulescu[2]

[1] *Institute for Mech. Engineering Problems and University of Technology and Design, Saint Petersburg, Russia.* , vakulenfr@mail.ru

[2] *DIMNP - UMR 5235 CNRS/UM2/UM1, Place E.Bataillon, Université de Montpellier 2, CP 107, 34095 Montpellier Cedex 5, France*

Flexibility and robustness are important properties of living systems in general, most particularly observed during development from egg or embryo to a fully organized organism. Flexibility means the capacity to change, when environmental conditions change. It allows the system to accommodate large perturbations. From a dynamical point of view we expect that flexible systems can support complicated attractors, periodic, chaotic as well as coexistence of several steady states (multi-stationarity). Somehow opposite, robustness is the capacity to support homeostasis in spite of perturbations. Intriguingly, biological systems are in the same time robust and flexible. Development of an organism is robust to variations of initial conditions and environment. Moreover, it is conceivable that intermediate developmental processes are flexible but lead to a reliable final result.

Centralized or bowtie architectures have been proposed as archetypes for robust intracellular networks (Aldana and Cluzel 2003; Ma et al 2007). We extend and prove rigorously similar ideas, valid for a more general class of dynamical models. Robustness and flexibility are discussed together within the same formalism.

We consider gene circuit models proposed by Reinitz, Mjolness and Sharp (1991, 1995). These models, inspired by the work of Hopfield (1982), were investigated intensively throughout the last decades in connection with segmentation of Drosophila (for example, Vakulenko et al. 2009). We study networks with a centralized architecture and show that this choice guarantees both flexibility and robustness. In particular, for these networks it is possible to program complex dynamical behavior by adjusting the interaction weights.

Such a network has K nodes of high output and input degree (centers), and $N >> K$ nodes of small degree (satellites). The control of the satellites by the centers is based on the divide and rule principle and assumes that there are no direct interactions between satellites. We also assume that centers are slow and that satellites are fast. This structure is possible in transcription factors – microRNAs networks, where the centers are transcription factors.

We show that these network can have $\exp(K \log N)$ stable steady states (rich multi-stationarity). Also, these network can exhibit, up to small corrections, arbitrary K-dimensional dynamics (if K=2, any kind of time periodic dynamics and for $K > 2$ chaotic dynamics). Such a

behavior is robust under large perturbations. For TF-microRNA networks we confirm, therefore, the idea that microRNAs increase robustness of patterning processes.

These networks can also be used as toy mathematical models for understading how positional information can be transformed into complex spatio-temporal patterns. One can describe how, smooth regular morphogen gradients create multicellular systems consisting of many cells, where each cell can have complicated dynamics.

This study raises some interesting open questions as follows. We studied centralized networks in a regime of total control, when satellite concentrations are exact functions of the center state. We showed that total control prevents a network from being robust, flexible and fast simultaneously, due to fundamental restrictions to the system parameters. An alternative to total control is a 'restricted' democracy, when there are possible fluctuations in the satellites concentrations. The open question is whether the restricted democracy releases some of the above constraints.

References:

1. M . Aldana and P.Cluzel, (2003) A natural class of robust networks, Proc. Natl. Acad. Sci. U. S. A., 100, 8710.
2. H. Ma, A. Sorokin, A Mazein, A. Selkov, E. Selkov, O. Demin, and I. Goryanin (2007), The Edinburgh human metabolic network reconstruction and its functional analysis, Molecular Systems Biology, 3, 135.
3. E. Mjolness,D. H. Sharp and J. Reinitz (1991), A connectionist Model of Development, J. Theor. Biol. V152, pp. 429-453.
4. J. Reinitz and D. H. Sharp (1995), Mechanism of formation of eve stripes, Mechanisms of Development, V49, 133-158.
5. J. J. Hopfield (1982), Neural networks and physical systems with emergent collective computational abilities, Proc. of Natl. Acad. USA}, V79, 2554-2558.
6. S. Vakulenko, Manu, J. Reinitz, O.Radulescu (2009), Size Regulation in the Segmentation of Drosophila: Interacting Interfaces between Localized Domains of Gene Expression Ensure Robust Spatial Patterning, Phys. Review Letters, V103, pp. 168102 - 168106

Detection of the set of features for discrimination of normal and high grade cancer tissues based on the Affymetrix microarray expression data.

Anna Karyagina[1,2,3], Anna Ershova[1,2,3], Michail Vasiliev[1,4], Ilya Lossev[5]

[1]*N.F. Gamaleya Research Institute of Epidemiology and Microbiology, Moscow, Russia*
[2]*Institute of Agricultural Biotechnology, Moscow, Russia*
[3]*A.N. Belozersky Institute of Physical and Chemical Biology, Moscow State University, Moscow, Russia*
[4]*Moscow Institute of Physics and Technology, Dolgoprudny, Moscow Region, Russia*
[5]*Parascript LLC, 6273 Monarch Park Place Longmont, CO 80503 USA*
mickvav@gmail.com

Motivation and Aim. Standard Affymetrix technology evaluates gene expression by measuring the intensity of mRNA hybridization with 25-mer oligonucleotide probes; the probes are grouped into probe sets. The probes within a probe set may work very differently; standard technique uses the 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. This approach was successfully implemented for the model of gender discrimination (Karyagina et al., 2010). It is based on selection of the features like "expression of a probe is greater (less) than some threshold" and construction of very simple, but effectively working recognizer. Now the similar approach with slightly modified feature selection algorithm is applied to discriminate high grade cancer and normal tissues. Our purpose was to reveal the set of features allowing to separate high grade cancer from normal tissues and to analyze the obtained high grade cancer expression profiles in different tissues.

Materials and Methods. CEL files obtained using Affymetrix 133 A, 133 Plus, 133 A_2 or 133 Plus_2 Arrays for different cancer and normal tissues were extracted from Array Express and GEO repositories. The data only for grade 3 and 4 cancer samples were considered. For training (feature selection) 391 CEL files of 5 normal and 234 CEL files of 4 cancer tissues were used. For testing 1484 and 541 CEL files of 44 normal and 28 cancer tissues were used, which are not included in training. The normalized PM values for each probe were used.

We consider two types of binary features: (a) the statement "expression of a probe is greater than threshold" and (b) "expression of a probe is less than threshold". Then we select the features of type *a* and *b* that are true with high frequency on cancer and small frequency on norm (cancer features) and features that are true with high frequency on norm and small frequency on cancer (norm features). One probe can belong simultaneously to both types of features. This means that large value for this probe is evidence for cancer and, simultaneously, small value is

evidence for norm, or *vice versa*. For each sample we compute number of cancer features that are true for this sample (number of cancer votes) and number of norm features that are true for this sample (number of norm votes). We consider sample to be cancerous if and only if number of cancer votes is larger than number of norm votes.

Results. Several sets of features for discrimination of normal and high grade cancer tissues based on the Affymetrix microarray expression data were selected using varied criteria of selection. In all cases the obtained recognizer included the probes of very similar lists of genes. The intersection of these lists included 8 main genes: three of them, AURKA, BOLA2 and DBF4 participate in cell cycle regulation, NUP37 takes part in mitotic progression, ECT2 participates in regulation of cytokinesis, CRY2 is a circadian regulator, mitochondrial gene MAOB codes monoamine oxidase, PPAP2B is essential for cell adhesion and migration. According to literature data most of these genes except for BOLA2 and NUP37 are associated with cancer.

For test data sets correct recognition of normal and cancer samples was observed in 95.3 and 90.0% of cases correspondingly. Two types of normal tissues: colon and stomach and two types of cancer tissues: prostate and kidney provide most part of errors (due to different explicable reasons) and thus this recognition method is not applicable to them. After excluding colon and stomach tissues recognition rate for normal tissues was 98.2%. After excluding prostate and kidney tissue recognition rate for cancer was 95.0%. The obtained data may be useful for understanding the common basis of cancerogenesis, as well as for development of simple universal express test-systems (for example, based on RT-PCR) for analysis of tissue samples from cancer patients.

1. A.Karyagina et al. (2010) Probe-level Universal Search (PLUS) algorithm for gender differentiation in Affymetrix data sets, *Journal of Bioinformatics and Computational Biology*, **8**:553–577.

Oligomerization of Bax in the Mitocondrial Outer Membrane upon Apoptosis

Valery Veresov, Alexander Davidovskii

Institute of Biophysics and Cell Engineering, National Academy of Sciences of Belarus, Belarus,
veresov@ibp.org.by

Interactions of Bcl-2 family proteins regulate permeability of the mitochondrial outer membrane upon apoptosis. Bax, a pro-apoptotic Bcl-2 family protein, translocates to mitochondria during apoptosis, where it brings about mitochondrial outer membrane permeabilisation (MOMP). MOMP releases proapoptotic factors, such as cytochrome c and SMAC/Diablo, into the cytosol where they activate caspases that leads eventually to cell destruction and apoptosis. Bax is a monomeric protein in the cytosol and integrates into the MOM only after activation by BH3-only proteins or other factors (1). Active Bax in the MOM forms oligomers that are required for MOMP (1). To understand membrane permeabilization by Bax, it is important to characterize the interface between the proteins in the oligomers. Recently two interfaces were suggested based on site-specific photocross-linking with the use of the detergent Triton X-100 as membrane surrogate (2). However, protein conformation within the membrane depends heavily on the membrane geometry and surface charge which cannot be reproduced adequately by this detergent. Here, the configuration of Bax monomer within the MOM obtained earlier with the use of an implicit membrane model (3) was refined using long run (200 000 full Monte Carlo- minimization (MCM) cycles) on the computer. Then the refined structure was applied to simulate interactions between monomers and to establish the interfaces. A four-stage simulation protocol with implicit membrane model (3) was used. The final configuration from (3) was taken as the starting one. At the first stage, the dihedral angles of the protein except those of the backbones of α-helices were taken as free. External rigid body translational and rotational motions were also allowed. At the second stage all dihedral angles were free to rotate. At the third stage, the computer molecular docking using the software PIPER (4) was used to model the dimerization of monomers. At the fourth stage, the dimer translational and rotational motions were allowed, while changes of all dihedral angles were forbidden. The data on the lipid composition of the outer leaflet of the MOM contact sites were used for the calculation of the effective surface charge density σ of the model membrane, which was assumed to be the sum of negative surface charge density due to acidic lipids in the membrane and a positive surface charge density due to membrane-adsorbed cations. Monte-Carlo-minimization strategy (5) with ECEPP2/ ECEPP3 protein force-field parameterization (Dunfield et al. 1978, Nemethy et al. 1983, Nemethy et al. 1992) was used at the first two stages. Solvatation of

protein residues was taken into account by the program GETAREA. The free energies of partitioning of amino acids from water into the cell membrane were taken into account by the hybrid Kessel-Ben-Tal (KBT) – White-Wimley (WW) –scale (KBT-LysWW-scale (3)) where the value of 7.4 kcal/mol for Lys of KBT – scale was replaced by that of 2.71 kcal/mol from the WW- scale. Two minimum free energy configurations of Bax-dimer at the MOM are shown in Fig.1.

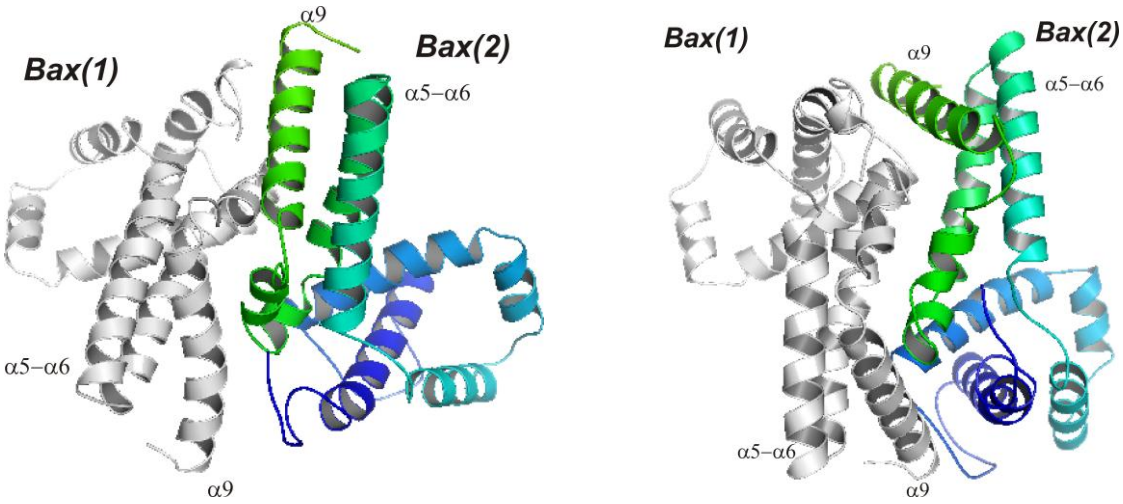


Fig.1. Two models of Bax-dimers.

The results suggest that the Bax-dimer, and likely Bax-oligomer, is formed by the insertion of helix α9 of one molecule into the groove formed by helices α9, α5 and α6 of the other molecule. These binding interfaces were not found in site-specific photocross-linking with the use of the detergent Triton X-100 as membrane surrogate (2) that can be explained that the configuration (conformation and orientation relative to membrane) within this detergent is different from that in MOM

1. R. J. Youle, A. Strasser (2008) The BCL-2 protein family: opposing activities that mediate cell death, *Nature Rev Mol Cell Biol*, **9**: 47-59
2. Z. Zhang‡, W. Zhu, S. M. Lapolla, D. W. Andrews, J. Ling, et al (2010) Bax forms an oligomer via separate, yet interdependent, surfaces, *J Biol Chem*, **285**, 17614–17627
3. V.G. Veresov, A.I. Davidovskii (2009). Activation of Bax by joint action of tBid and mitochondrial outer membrane: Monte Carlo simulations, *Eur Biophys J*, **38**: 941-960
4. D. Kozakov, R. Brenke, S. R. Comeau, S. Vajda (2006) PIPER: An FFT-based protein docking program with pairwise potentials, *Proteins*, **65**:392–406
5. Z. Li, H. A. Scheraga (1987), Monte Carlo – minimization approach to the multiple-minima problem in protein folding, *Proc Natl Acad Sci USA*, **84**, 6611-6615

Integration of the Protein Bcl-2 into Mitocondrial Outer Membrane upon Apoptosis

Valery Veresov, Alexander Davidovskii

Institute of Biophysics and Cell Engineering, National Academy of Sciences of Belarus, Belarus,
veresov@ibp.org.by

The antiapoptotic protein Bcl-2 inhibits apoptosis by antagonizing proapoptotic Bcl-2 family members (1). Bcl-2 is constitutively bound either to mitochondrial outer membrane (MOM) or endoplasmic reticulum with helix 9 embedded in the bilayer and the rest of polypeptide located on the cytoplasmic side of the membrane (2). Despite more than fifteen years of analysis and a plethora of experimentally tested models, the molecular details of the mechanism by which Bcl-2 prevents apoptosis remain elusive. Experimental data show that Bcl-2 undergoes a dramatic conformational change during apoptosis after interaction with tBid resulting in the insertion of the hairpin α5-α6 from an aqueous to hydrophobic environment (2-4). However how this occurs at molecular level still remains uncertain. Here, the mechanisms of integration of Bcl-2 into a model membrane mimicking the MOM were studied by Monte Carlo simulations preceded by a computer prediction of the docking of tBid with Bcl-2. Experimentally based structural constraints as well as physical and evolutionary considerations were used to reduce the conformational space to be searched. For this purpose, the process of integration of Bcl-2 into the MOM was divided, on the basis of a number of experimental data, into four stages with four different procedures for conformational space reductions and with four types of Monte Carlo simulations. These four stages were: (i) the insertion of helix α9 into the membrane; (ii) the formation of the complex Bcl-2-tBid; (iii) the rotation of the complex as a whole rigid body due the interactions with the anionic lipid charges of MOM; (iv) the insertion of hairpin α5-α6 into the membrane. At the first stage the dihedral angles except those of the residues within the loop α8-α9 were frozen while translations and rotations of Bcl-2 as a whole rigid body were permitted. At the second stage the software PIPER (6) was used to model the docking between Bcl-2 and tBid. At the third stage only translations and rotations of the complex as a whole were allowed. At the fourth stage only the dihedrals from the loops α4-α5 and α6-α7 were allowed to be changed at first followed by a refinement procedure with all dihedrals except those of the backbones of α-helices taken as free to rotate. At the third stage all dihedral angles were taken as free to change. Monte-Carlo-minimization strategy (7) with ECEPP2/ ECEPP3 protein force-field parameterization was used at the stages I and IV. The solvation of protein

residues was taken into account by the program GETAREA. The free energies of partitioning of amino acids from water into the cell membrane were taken into account by the hybrid Kessel-Ben-Tal (KBT) – White-Wimley (WW) –scale (KBT-LysWW-scale (5)).

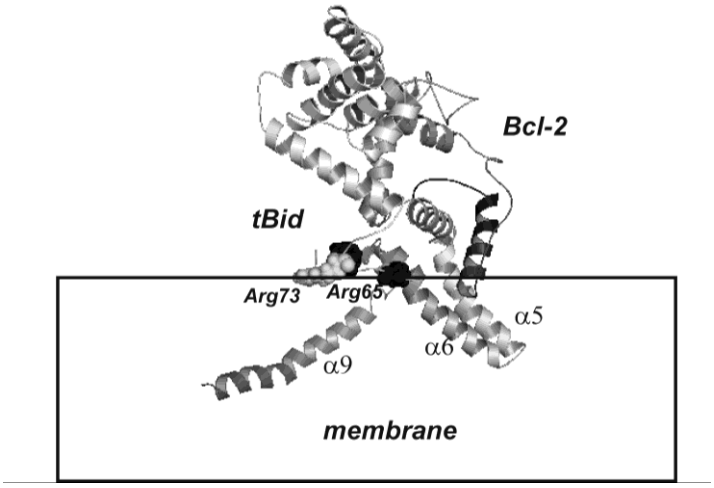


Fig.1. A model of Bcl-2 integrated into the MOM.

A novel model of Bcl-2 integration into the MOM activated by interaction with tBid was predicted by the simulations. In this model, tBid binds to Bcl-2 at an interaction site formed by Bcl-2 near helix α1 leading, due to interaction of the positively charged N-terminal fragment of tBid with anionic lipid headgroups, to Bcl-2 reorientation such that the hairpin α5-α6 is brought into close proximity with negatively charged lipid headgroups followed by insertion of the hairpin into MOM (shown in Fig.1).

1. R. J. Youle, A. Strasser (2008) The BCL-2 protein family: opposing activities that mediate cell death, *Nature Rev Mol Cell Biol*, **9**: 47-59
2. P. K. Kim, M. G. Annis, D. W. Andrews et al. (2004) During apoptosis Bcl-2 changes membrane topology at both the endoplasmic reticulum and mitochondria, *Mol. Cell*, **14**: 523–529.
3. P. J. Dlugosz, L. P. Billen, Annis M. G., W. Zhu, Zhang Z., Lin J., B. Leber, D. W. Andrews. (2006) Bcl-2 changes conformation to inhibit Bax oligomerization, *EMBO J.* **25**: 2287–2296.
4. J. Peng, D. Andrews, J. Lin, et al. (2006) tBid Elicits a Conformational Alteration in Membrane -bound Bcl-2 Such That It Inhibits Bax Pore Formation, *J. Biol. Chem.*, **281**: 35802–35811
5. V.G. Veresov, A.I. Davidovskii (2009). Activation of Bax by joint action of tBid and mitochondrial outer membrane: Monte Carlo simulations, *Eur Biophys J*, **38**: 941-960
6. D. Kozakov, R. Brenke, S. R. Comeau, S. Vajda (2006) PIPER: An FFT-based protein docking program with pairwise potentials, *Proteins*, **65**:392–406
7. Z. Li, H. A. Scheraga (1987), Monte Carlo – minimization approach to the multiple-minima problem in protein folding, *Proc Natl Acad Sci USA*, **84**: 6611-6615

Identification of amino acid residues defining substrate specifity of cytochrome P450

Alexander Veselovsky, Maria Zharkova, Dmitiriy Filimonov, Boris Sobolev
Institute of Biomedical Chemistry RAMS, Russian Federation, veselov@ibmh.msk.su

The cytochrome P450 superfamily is a large group of heme containing monooxygenases exhibiting various functions in the cells. Several members of this superfamily are the main enzymes of xenobiotics metabolism, particular the drugs. Although this family is characterized by a very conservative spatial structure of the protein, members of this family may reveal the weak similarity in amino acid sequences. The present classification of cytochrome P450 superfamily is based on similarity of amino acid sequences and does not correlate with substrate specifity of enzymes.

At present several tools have been developed to prediction of drug metabolism by cytochrome P450 based on structures of ligands. These methods suggest that all information needed for prediction are included in structure of ligands, but in many cases such predictions are unsuccessful. One of the reasons is inconsiderable of structure of the active sites of cytochrome P450 and amino acid residues participating in recognition of substrates.

The recent results in identification of amino acid residues defining substrate specifity of cytochrome P450 will be discussed.

Molecular phylogenetic approach to study of the earliest land plants: the family Cephaloziaceae Mig. s.l. (Marchantiophyta)

Anna Vilnet[1], Nadezda Konstantinova[1], Alexey Troitsky[2]
[1]*Polar-alpine Botanical Garden-Institute of Kola SC RAS, Russian Federation, anya_v@list.ru*
[2]*A. N. Belozersky Institute of Physicochemical Biology of Moscow State University, Russian Federation*

During last decade molecular phylogenetic studies proved that liverworts appear to be the earliest diverged phyla of higher plants. The liverworts together with mosses and hornworts for a long time had been united in one phylum based on dominating of haploid phase in a life cycle. Molecular phylogenetic studies resolved these groups in three separated phyla. In general backbone phylogeny of liverworts is clarified and modern classification of the group is supposed. The small plants size, insufficient studied variability and limited number of features together with a wide distribution and ecological plasticity of majority of species led to

difficulties in study of liverworts taxonomy. Recently molecular studies of liverworts focus mostly at families and genera level.

The family of leafy liverworts Cephaloziaceae s.l. is world widely distributed. There are different treatments of Cephaloziaceae. Some systematics segregated families Odontoshismataceae (Grolle) Schljakov, Hygrobiellaceae Müll. Frib from Cephaloziaceae but in recent classification of Crandall-Stotler et al. (2009) based on molecular data Cephaloziaceae treated as one complex family with 16 genera. We attempt to clarified a phylogeny of Cephaloziaceae based on ITS1-2 nrDNA and trnL-F cpDNA sequence data analyses of species from genera Cephalozia (Dumort.) Dumort., Odontoschisma (Dumort.) Dumort., Nowellia Mitt., Pleurocladula Grolle, Schofieldia Godfrey, Hygrobiella Spruce, Cladopodiella H. Buch, Alobielopsis R.M. Schust. and Iwatsukia N. Kitag. We revealed that Cephaloziaceae s.l. is not monophyletic: particularly Odontoschismataceae and Cephaloziaceae s.str. are in a sister relation, whereas Hygrobiellaceae clusters in a clade with unclear affinities. We suppose that treatment of Odontoschismataceae and Hygrobiellaceae as distinct families is more natural.

The genera Cephalozia, Odontoschisma and Cladopodiella are not monophyletic as well. The monotypic genera Pleurocladula and Schofieldia, also as Nowellia are located in a Cephalozia-clade. The species of Cladopodiella and Odontoschisma are intermingled in one clade. The taxonomical rearrangements at genera level in Cephaloziaceae and Odontoschismataceae should be implemented with careful reevaluation of morphological data. Several taxa that treated previously as infraspecific level appear to be a separate species. Cephalozia otaruensis Steph. was accepted earlier as subspecies of Cephalozia bicuspidata (L.) Dumort. now is shown as phylogenetically allied but clearly separated species. Cephalozia affinis Lindb. ex Stephani molecularly quit distinct from C. lunulifolia with that it was earlier synonimized by many authors. The high level of infraspecific heterogeneity was found for Cephalozia pleniceps (Austin) Lindb., possibly some new taxa could be described from this complex. The variability in species of Odontoschisma and Cladopodiella are very low. The score of divergent events in sister lineages is different. Suboceanic Hygrobiella laxifolia (Hook.) Spruce from monotypic genus possibly is presented by two species: one of them occurred in Northwest of European Russia and other – in Far East.

This work was partially supported by Russian Foundation for Basic Researches grants (09-04-00281, 09-04-01324, 10-04-00050) and President Program for support of PhD researches (MK-3328.2011.4).

Genome-wide analysis of possible ncRNA structures

Svetlana Vinogradova, Andrey A. Mironov
Moscow State University, Russian Federation, kintany@gmail.com

Non-coding RNAs (ncRNAs) are functional transcripts that do not encode proteins. Recent discoveries shown that the diversity and importance of ncRNAs were underestimated [1]. However, in contrast to protein coding genes, the signals for ncRNA are subtler. The one general characteristic shared by many (but not all) known RNAs is folding into complex shapes that are crucial to function and thus are conserved.

It is possible to predict RNA structures by similarity and there are several possibilities to do this. The first approach uses primary sequence information to align the query sequence to the target database. Such searches are exemplified by programs like BLAST and FASTA. These sequence alignment programs are $O(N^2)$ in time and memory, where N is the length of the sequences being analyzed. The second one consists of a search with a known RNA structure against a sequence database. Such searches have been implemented with profile stochastic context free grammars (SCFGs) and require $O(N^3)$ memory and $O(N^4)$ time. However, computational *de novo* discovery of ncRNAs within genomic sequences is still in its infancy.

Although it is obvious that searching for possible structured elements in one nucleotide sequence is generally not very effective, searching in multiple (orthologous) sequences can be highly effective, since evolutionary conservation highlights functionally important regions of all kinds. Importantly, such searches leverage the rapidly increasing body of comparative genomic sequence data. A key issue, however, is that the evolutionary signature of an RNA gene is quite different from that of a protein-coding gene. So fast and effective algorithm analyzing multiple alignments and searching for conserved RNA structures could be a powerful tool.

We have developed an algorithm for genome-wide prediction of possible RNA structures. It is based on computing of local RNA base pairing probabilities [2] for sequences in multiple alignments and their further analysis in terms of maximums or sums in a sliding window. The efficiency of the algorithm was tested on full alignments of 14 insects with D. melanogaster, and it has been shown that it allows to efficiently retrieve candidates for microRNA precursors or other structured RNAs.

1. A. Huttenhofer, P. Schattner, N. Polacek (2005) "Non-coding RNAs: hope or hyp" *Trends Genet.* **21(5)**: 289–297
2. S. H. Bernhart, I.L. Hofacker, P.F. Stadler (2006) "Local Base Pairing Probabilities in Large RNAs" *Bioinformatics*, **22:** 614-615

Improved prediction of human miRNAs based on HMMs and the analysis of "young" miRNAs

Pavel Vorozheykin

Novosibirsk State University, Russian Federation, pavel.vorozheykin@gmail.com

MiRNAs are a large family of small non-coding RNAs that control mRNA expression either by the cleavage or by the translation arrest. MiRNAs are defined as single-stranded RNAs of 19-25 nt in length processed from miRNA precursors (pre-miRNAs) that form stem-loop secondary structure.

Computational prediction of miRNAs and their precursors is a constantly growing topic of bioinformatics and while miRNAs functioning becomes increasingly better investigated origin and evolution of miRNAs remain largely obscure.

In this report we present HMM-based methods for the prediction potential human miRNAs and their precursors using statistical information about known sequence data.

To find distant miRNAs as well as close homologs Nam with coauthors [1] have introduced a probabilistic method based on a paired hidden Markov model (HMM) for miRNA genes which simultaneously considers the structure and sequence of pre-miRNAs. ProMiR is preformed well on the first pre-miRNA datasets, but we found that it frequently fails to detect miRNA for recently discovered miRNA even after re-learning on the last datasets.

To predict human pre-miRNAs we improved new hidden Markov models which include context-structural characteristics of ProMiR [1] model, but exclude excess of model parameters. Using the 12.0 MIRbase release the comparison of the efficiency of human pre-miRNA predictions shows that the sensitivity of our method is higher (0.90) than ProMiR (0.73) with the same specificity (0.96). False positive and false negative errors of our method are more suitable than ProMiR's. Also to increase the accuracy of human miRNA prediction the HMM algorithms of Nam et al. [1] were modified by taking into account the loop frequency distribution. It shows about 2 nt more accurate miRNA boundary.

Scanning human genome for the known miRNAs we find a surprisingly huge number of 2122 homologs which may be considered as "young" or "candidate" miRNAs. Identifying the localization of the miRNA copies within the mobile elements we develop the model of miRNA duplication. We have found that either miRNA or young miRNA nearest-neibor-distance distributions obey the similar Weibull laws. We have performed the simulations of miRNA inverted duplications and have fitted the Weibull parameters to the observed miRNA genomic distribution.

The comparison between the secondary structures of the original miRNAs and their copies suggests the evolutionary selection acting on the miRNA expansion.

1. J. Nam et al. (2005) Human microRNA prediction through a probabilistic co-learning model of sequence and structure, Nucleic Acids Research, 33(11):3570-3581.

Co-regulating miRNA clusters are functionally conserved but widely dispersed in protein-protein interaction networks

Wilson Goh[1,3], Kwok Pui Choi[2] and Limsoon Wong[3*]

[1] Department of Computing, Imperial College London, UK; [2] Department of Statistics and Applied Probability, National University of Singapore, Singapore; [3] School of Computing, National University of Singapore, Singapore

We implemented a statistical pipeline to uncover co-regulating miRNA clusters, that is, groups of miRNAs that target shared genes. Although these clusters are functionally conserved, we found this functional conservation does not translate to physical proximity in gene networks. This suggests that co-operating miRNAs achieve silencing not by acting on a single pathway or complex, but by mass action across a wide repertoire of targets that are involved in related functions. In addition, we also find that most co-regulating clusters are co-expressing suggesting that cooperation rather than substitution between co-regulators is the norm. However, a small subset of co-regulating clusters possesses a single anti-co-expressed member. We postulate that the anti-co-expressed member may help to fine tune the functionality of the cluster under specific expression conditions.

Using Diana[1], we derived individual miRNA-target precisions from the signal-to-noise ratio. For every miRNA pair, we calculated a hypergeometric p-value based on their target overlap (1000 simulations, $p<0.01$). Significantly co-regulating miRNA pairs were modeled as a network from which co-regulating clusters were extracted as maximal cliques. The gene targets in these miRNA clusters had to be shared by over 90% of members. miRNA cluster target genes were assigned Gene Ontology (GO) terms based on hypergeometric test ($p <= 0.01$). To limit GO terms to only informative terms, we implemented a GO term cut-off (at least 30 genes annotated to the term, no child term > 30 genes). To evaluate cluster GO term coherence, we calculated average path lengths (P_{avg}) between all GO terms annotated to cluster. We randomly generated clusters with equal informative GO terms and calculated P_{avg}1000 times to convert cluster P_{avg} into a z-score. We combined protein-protein interaction information from HPRD, BioGrid, IntAct, DIP, MINT, and literature. The integrated dataset was then cleaned using the methods described by Liu[2] and integrated with Pathway API[3]. miRNA cluster targets were mapped onto the combined network and BioGrid respectively. A z-score is derived from cluster average path length analysis by randomly selecting a set of nodes and calculating the average pathlength 1000 times. Links between miRNA expression pairs are profiled using methods described by Wu *et al* [4]. We modeled correlation between miRNA pairs as Hamming Distance, D. Cluster expression coherence is the average of the D calculated for all member miRNA pairs.

129 miRNA clusters were derived from co-regulation network. Using Tarbase, our co-regulation network has nine fold enrichment of real targeting relations using log odds ratio. It

was also found the inter-Biological Process GO term distances were significantly smaller, with more than 70% having a z-score lower than -1.96. Interestingly, congruent GO term annotation does not translate to clustered nodes in a gene interaction network. We calculated z-scores based on average path lengths in both BioGrid and an integrated network. In both cases, cluster targets are not proximal on the gene interaction network.

Most miRNA co-regulation clusters are also expression-ally similar. Most clusters have an average D of 0 or near to 0. This is expected considering that many intra-cluster miRNAs belong to the same family. However, a subset of 30 co-regulation clusters contains a miRNA whose D is very high. To better understand the role of such miRNAs, we removed them from the cluster and checked if it was accompanied by a rise in cluster target membership. In 50% of cases, there were small change in gene memberships. In the rest, a big jump (~50%) was observed. This suggests that the anti-co-expressed component targets many genes not co-shared with the other members. Where a gene membership jump was observed, we computed the change in informative GO term distance. Where a large z-score was previously observed, removal of the anti-co-expressed component consistently lowered this value. This implies these clusters are functionally coherent as they are co-expressed normally without the anti-co-expressed component.

1. Maragkakis, M. et al, *Nucleic Acids Res* **2009,** 37, (Web Server issue), W273-6.
2. Liu, G. et al,.*Bioinformatics* **2009,** 25, (15), 1891-7.
3. Soh, D. et al.. *BMC Bioinformatics* **2010,** 11, 449.
4. Wu, J. et al, *Bioinformatics* **2003,** 19, (12), 1524-30.

Putative Causes of the Disagreement between the Transcription Rates of Transposable elements and Their Transposition Frequency in y cn bw sp Strain of Drosophila melanogaster

Lyudmila Zakharenko[1], Tatyana Bak[2], Olesya Ignatenko[2]

[1]*Institute of Cytology and Genetics, Siberian Branch of the Russian Academy of Sciences, pr. Lavrent'eva, 10, Novosibirsk, 630090 Russia, zakharlp@bionet.nsc.ru*
[2]*Novosibirsk State University. Pirogova, 2, Novosibirsk, 630090 Russia*

Transposable elements (TEs) are present in all genomes investigated hitherto. They can migrate over the genome. Transposable elements constitute about one-quarter of the Drosophila melanogaster genome. They belong to nearly 100 families, and each family includes tens of copies on the average. Some copies are full-size and able to produce functional enzymes. However, the majority of TE copies are truncated and unable to induce transposition. Some truncated copies can migrate owing to the presence of terminal repeats and cis recombinations. Probably, this is one of the causes of the fact that no direct correlation between the number of TEs in the Drosophila genome and the rate of their transposition has been detected.

We investigated the effect of TE transcription rate on their transposition rate. The rate of TE transposition (mdg1, mdg2, copia, roo, hobo, I-element) was assessed from the difference between the current TE locations in the genome of Drosophila melanogaster strain y cn bw sp, completely sequenced by now, and their locations according to in silico data obtained several years ago. Transposable elements can express in various tissues at various developmental stages, but only their expression in generative tissues can leave marks in future generations. According to the EST data by Deloger et al. [1], most TEs are not expressed in testes, whereas ovaries demonstrate high levels of activity of some TEs. We found no direct correlation between the expression rates of TEs in generative tissues of y cn bw sp and the TE transposition rate in this genome. For example, the hobo element intensely migrates over the genome with the transposition frequency 102 per site per genome per generation [2], whereas ESTs for hobo are not found in any organ [1].

Another objective of our study was the analysis of full-size copies of some TEs for the presence of open reading frames (ORFs) and the assessment of the significance of single-nucleotide substitutions for translation. For example, the high expression rate of mdg2 was not accompanied by a high transposition rate. We showed that most (23 of 24) full-size mdg2 copies

had open reading frames with lengths approximately corresponding to that of the annotated sequence. However, a large portion of the ORFs had single-nucleotide substitutions and inserts, which could alter the amino acid sequences of the corresponding polypeptides. As reported by Kudla et al. [3], synonymous substitutions do not alter the amino acid sequence of the protein but may affect the expression rate of the gene at the level of mRNA production and the rates of RNA and protein degradation. In this way, they influence the growth rate of Escherichia coli. The stability of mRNA folding near the ribosomal binding site plays the key role in this process. The mdg2 ORFs are polymorphic for amino acid sequences. The largest group includes five ORFs. Thus, the polymorphism of full-size TE copies at the levels of nucleotide and amino acid sequences may affect the function of the enzymes responsible for TE migration.

Acknowledgments: This work was supported by grant 09-04-00213a from the Russian Foundation for Basic Research, and grants B 27.29 and B 27.30 from the Biodiversity program of the Presidium of the Russian Academy of Sciences.

Literature

1. M. Deloger et al. (2009) Identification of expressed transposable element insertions in the sequenced genome of Drosophila melanogaster, Gene,. 439:55-62.
2. L. Zakharenko et al. (2007) Fluorescence in situ hybridization analysis of hobo, mdg1 and Dm412 transposable elements reveals genomic instability following the Drosophila melanogaster genome sequencing., Heredity99:525-30.
3. G. Kudla et al. (2009) Coding-Sequence Determinants of Gene Expression in Escherichia coli, Science, 324:255-258.

Structural Classification of Bacterial Serine/Threonine Protein Kinases

N. Zakharevich[1], D. Osolodkin[2], I. Artamonova[1], V. Palyulin[2], N. Zefirov[2], V. Danilenko[1]

[1]*Vavilov Institute of General Genetics, 119991, Moscow, Russia, Gubkina St., 3, zakharevich@yandex.ru*
[2]*Department of Chemistry, Moscow State University, 119991, Moscow, Russia,Leninskie Gory 1/3, dmitry_o@org.chem.msu.su*

A classification of eukaryotic-like serine-threonine protein kinases (ESTPK) of various gram-positive bacteria was proposed. Three groups of bacteria were considered (471 sequences total): pathogenic (*Actinomyces, Bacillus, Corynebacterium, Frankia, Nocardia, Mycobacterium, Staphylococcus, Streptococcus*), probiotic (*Bifidobacterium, Clostridium, Lactobacillus*) and kinase inhibitors producents (*Rhodococcus, Saccharopolyspora, Streptomyces, Thermobifida*).

ATP binding sites of selected kinases were compared with homologous regions of the bacterial kinases with known crystal structure (PknB, PknE and PknG *Mycobacterium tuberculosis*), in order to clarify the structural criteria of selectivity of potential ATP-competitive inhibitors. Absolutely conservative residues and residues contributing to binding only by backbone atoms were eliminated from consideration. Nine remaining residues represent specific signature of the adenine binding pocket (ABP). The kinases were grouped according to physico-chemical properties of the residues forming signatures. Gene conservation was not taken into account.

We identified 20 groups containing more than one kinase based on the ABP amino acid composition. The main criterion for group establishment was the presence of specific combinations of the hydrogen bond donors/acceptors, aromatic and hydrophobic residues in the ABP signature.

Substrate binding region is not very conserved due to necessity to bind various substrates, but ABP is also less conserved compared to the phosphate transferring region. This fact opens the possibility to explore this variability during the inhibitor design. Possible selectivity issues were studied with the help of homology modeling of the structure of typical representatives of each group.

The groups obtained are rather diverse, but a number of trends does exist. First of all, the ‘ceiling’ of the ATP binding site is usually hydrophobic. The first residue of all signatures is always a bulky hydrophobic residue, usually Leu or Ile. The hinge region is more variable, but functional importance of the sidechains is lower in this place. In most cases Met residue appears in gatekeeper position; Thr, Ile and aromatic gatekeepers are also rather common. The ‘floor’ of the ABP is the most variable region built mainly by the amino acid sidechains. The properties of

the residues differ substantially: they can be bulky or small, hydrophobic or hydroxyl containing, or even aromatic.

The signature-based classification was compared with the phylogeny of studied kinases. It was found that selectivity between closely related branches of phylogenetic tree can be achieved.

Kinases characterised by the same signature are most probably inhibited by same molecules which can be used as universal antibacterial compounds. Design of the inhibitors of kinases belonging to the large groups containing both kinases of pathogenic bacteria and kinases of healthful probiotic bacteria should be studied in more detail, in order to achieve selectivity. Kinases of the probiotic organisms should be considered as antitarget: their inhibition should be avoided.

Dynamic model of anaerobic energy metabolism of yeast *Saccharomyces cerevisiae*

Maksim Zakhartsev[1], Alexej Lapin[2], Matthias Reuss[2]

1IPMB, University of Heidelberg, Im Neuenheimer Feld 364, 69120 Heidelberg, Germany, zakhartsev@uni-heidelberg.de
2Center Systems Biology (CSB), University of Stuttgart, Nobelstr. 15, 70569 Stuttgart, Germany, lapin@ibvt.uni-stuttgart.de , reuss@ibvt.uni-stuttgart.de

Cellular energy metabolism, besides conversion of energy flow for metabolic purposes, is a metabolic hub that interfaces metabolic modules through which a metabolic perturbation can propagate from one module to another, thus implementing a signaling role. At steady state conditions, an adenylate pool (AXP = ATP + ADP + AMP) usually is assumed operating under 'conserved moiety' law. However, fast metabolic perturbation experiments (e.g. glucose-pulse) have revealed inconsistency of this assumption on a minute timescale, where AXP pool operates as an 'opened pool' as it was theoretically described in [1]. As a result of the glucose-pulse the entire AXP pool shrinks in a matter of 30 seconds and replenishes only in 5 minutes (so-called 'ATP-paradox'). This phenomenon was observed long time ago [2] however only recently a molecular mechanism underlying this regulation was elucidated [3,4 and personal observations]. In course of the perturbation-induced transitions the excess of AMP is ousting into inosine and hypoxanthine via purine salvage reactions and then adenylate pool is replenished through *de novo* pathway as well as salvage reactions. Consequently the question has arisen: what metabolic meaning does the ATP-paradox have? To better understand this phenomenon we have performed

glucose pulse experiment on anaerobically growing yeast in chemostat. ODE-based model of major metabolic regulatory events in glycolysis, pentose-phosphate pathway, purine *de novo* synthesis, nucleotide salvage reactions, redox balance and biomass growth was developed. Transient concentrations of the extra- and intracellular metabolites were used for parameterization of the model. The distinctive feature of this model is that the AXP dynamically balanced as 'opened moiety' through purine *de novo* synthesis pathway, purine salvage reactions and biomass growth. The model explains dynamic behavior of all measured metabolites and predicts that the rate of purine *de novo* synthesis increases after the glucose pulse, which would result in peaking of all intermediates along linear purine *de novo* pathway. This event is one of coupling points between metabolic and genetic regulations. To our knowledge, transient increase of (S)AICAR intermediate from purine *de novo* synthesis pathway can explain earlier observation in [5] on change in whole genome expression profile after the glucose pulse through de-repression of *Bas1* and *Pho2* transcriptional factors, which coordinate upregulation of the purine biosynthesis, sulfur and phosphorus assimilation, methionine and adenine salvage pathways. Thus, stimulus-response methodology aided with mathematical modeling has allowed us better understand of functional meaning of 'ATP-paradox' as preparation of a cell for transition from substrate limited to unlimited growth and its coupling with genetic regulation.

1. J.Reich, E.Sel'kov (1981) Energy metabolism of the cell: a theoretical treaties, *Academic Press*, pp.345.
2. U. Theobald, W. Mailinger, M. Reuss, and M. Rizzi (1993) *In vivo* analysis of glucose-induced fast changes in yeast adenine nucleotide pool applying a rapid sampling technique, *Anal.Biochem.*, **214 (1):**31–37.
3. M. Loret, L. Pedersen, and J. Francois (2007) Revised procedures for yeast metabolites extraction: application to a glucose pulse to carbon-limited yeast cultures, which reveals a transient activation of the purine salvage pathway, *Yeast*, **24 (1):**47–60.
4. T. Walther, M. Novo, K. Roessger, F. Letisse, M. O. Loret, J. C. Portais, and J. Francois (2010) Control of ATP homeostasis during the respiro-fermentative transition in yeast. *Molecular Systems Biology*, **6:**344.
5. M. Kresnowati et al. (2006) When transcriptome meets metabolome: fast cellular responses of yeast to sudden relief of glucose limitation, *Molecular Systems Biology*, **2**:Article number: 49.

Comparative genomics of Arthropods

Evgeny M. Zdobnov

Department of Genetic Medicine and Development, University of Geneva Medical School, 1 rue Michel-Servet, Geneva 1211, Evgeny.Zdobnov@unige.ch

I wish to present our major findings from comparative studies of recently sequenced Arthropod genomes: the freshwater crustacean Daphnia [1], the human body louse [2], the parasitic wasp Nasonia [3], as well as Culex [4, 5] & Aides [6-8] mosquitos.

Daphnia pulex is only 200 megabases and claimed to contain at least 30,907 genes. The high gene count seems to be a consequence of an elevated rate of gene duplication and adaptations to ecological challenges.

The human body louse is an obligatory human parasite and is an important vector for diseases, including epidemic typhus, relapsing fever, and trench fever. It has the smallest known insect genome, spanning 108 Mb. Despite its status as an obligate parasite, it retains a remarkably complete "basal insect" repertoire of 10,773 protein-coding genes and 57 microRNAs. Representing hemimetabolous insects the genome of the body louse thus provides the most complete reference for studies of holometabolous insects. Compared with other insect genomes, the body louse genome contains significantly fewer genes associated with environmental sensing and response, including odorant and gustatory receptors and detoxifying enzymes.

Generally unknown to the public, the parasitic wasps kill pest insects. Harnessing their potential would thus be vastly preferable to chemical pesticides. Besides also offering pharmaceutically interesting venoms, the wasps could act as a new genetic system with a number of unique advantages. So far, fruit flies have been the standard model for genetic studies, mainly because they are small, can be grown easily in a laboratory, and reproduce quickly. On top of sharing these traits, Nasonia present another advantage. Male Nasonia have only one set of chromosomes. We also identified changes to metabolic pathways that may reflect the amino-acid rich carnivorous diet of these parasitoids.

Sequencing of the Culex mosquito genome achieves the important goal of obtaining a complete reference genome from each of these three major taxonomic groups of disease-vector mosquitoes, in addition to Anopheles gambiae and Aedes aegypti. We integrated the results from experimental expression analyses of 25 different vector-pathogen interactions with the computational comparative genomics data to reveal key mosquito genes responsive to diverse pathogen infections including viruses, filarial worms, bacteria, and malaria parasites. The expanded Culex immune-related gene repertoire uncovered evidence that naturally mosquito-borne pathogens have evolved to evade the vectors' innate immune responses while non-native mosquito-pathogen interactions cause systemic responses.

1. Colbourne, J.K., et al. (2011.) The ecoresponsive genome of Daphnia pulex. *Science*, **331**(6017): p. 555-61.
2. Kirkness, E.F., et al. (2010.) Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle. *Proceedings of the National Academy of Sciences of the United States of America*, **107**(27): p. 12168-73.
3. Werren, J.H., et al. (2010.) Functional and evolutionary insights from the genomes of three parasitoid Nasonia species. *Science*, **327**(5963): p. 343-8.
4. Bartholomay, L.C., et al. (2010.) Pathogenomics of Culex quinquefasciatus and meta-analysis of infection responses to diverse pathogens. *Science*, **330**(6000): p. 88-90.
5. Arensburger, P., et al. (2010.) Sequencing of Culex quinquefasciatus establishes a platform for mosquito comparative genomics. *Science*, **330**(6000): p. 86-8.
6. Waterhouse, R.M., S. Wyder, and E.M. Zdobnov (2008.) The Aedes aegypti genome: a comparative perspective. *Insect Molecular Biology*, **17**(1): p. 1-8.
7. Waterhouse, R.M., et al. (2007.) Evolutionary dynamics of immune-related genes and pathways in disease-vector mosquitoes. *Science*, **316**(5832): p. 1738-43.
8. Nene, V., et al. (2007.) Genome sequence of Aedes aegypti, a major arbovirus vector. *Science*, **316**(5832): p. 1718-23.

Regulation of Multidrug Resistance Genes by Transcriptional Factors from the MerR Family

Ilya ZHAROV[1,2], Mikhail GELFAND[2,3], Alexey KAZAKOV[2,4]

[1]*Moscow Institute of Physics and Technology, Moscow 117303, Russia*
[2]*Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia*
[3]*Department of Bioengineering and Bioinformatics, Moscow State University, Moscow, Russia*
[4]*Lawrence Berkeley National Laboratory, Berkeley, California, 94720, USA*
peshwalk@gmail.com, gelfand@iitp.ru, kazakov@iitp.ru

Development of the multidrug resistant (MDR) phenotype in bacteria is a major issue in healthcare. The MFS-family multidrug transporters Bmr and Blt were experimentally studied in *Bacillus subtilis* [1, 2, 3]. They export the same range of antibacterial cationic compounds from the cell using the proton antiport mechanism. Transcription of their genes is activated by the MerR-family transcriptional factors (TFs) BmrR and BltR, respectively [3,4]. These TFs have homologous N-terminal DNA-binding domains but non-homologous C-terminal ligand binding domains. Some of the Bmr substrates are confirmed BmrR ligands [4]. The BltR ligands have not been identified yet. In addition to deleterious compounds, Blt exports polyamine spermidine [5]. Its gene is cotranscribed with *bltD* gene that encodes spermine/spermidine acetyltransferase [6]. Acetylation of these polyamines leads to their degradation.

We studied binding sites and regulons for 45 orthologs of BltR and 152 orthologs of BmrR using a comparative genomic approach. These proteins are mainly found in the Firmicutes and the Actinobacteria. A phylogenetic tree was built for studied TFs using their N-terminal domains. To search for binding sites, we built positional weighted matrices (PWMs) based on

experimentally studied BmrR and BltR binding sites from *B. subtilis*. A selective PWM was built for each major branch of the TFs phylogenetic tree. Putative binding sites of these TFs are located in long (19−20 bp) spacers between the −35 and −10 promoter boxes of regulated operons. This arrangement is typical for the activators from the MerR family. The BltR and BmrR palindromic motifs overlap with the −35 promoter box. We used this feature to eliminate false positive sites found with PWMs. As only 1−3 true positive sites per TF were found, we suggest that studied TFs are local regulators. Divergently transcribed regulated operons are typical for BltR orthologs, whereas BmrR-regulated operons are usually transcribed in the same direction with TF-encoding genes.

Regulated operons for studied TFs usually comprise 1−2 genes. Most-frequently regulated genes are multidrug transporters of various groups: MFS (H+-antiporters), MATE (Na+/H+-antiporters) and ABC (ATP hydrolysis driven). Their role was proposed based on the phylogenetic analysis. Regulation of structurally dissimilar but functionally equivalent transporters by studied TFs confirms their role as MDR regulators. Another gene frequently regulated by BltR and BmrR orthogs encodes spermine/spermidine acetyltransferase homologous to BltD of *B. subtilis*. This finding demonstrates a link between multidrug resistance and polyamine metabolism in Gram-positive bacteria.

1. S. Grkovic, M.H. Brown, R.A. Skurray (2002) Regulation of bacterial drug export systems, *Microbiol. Mol. Biol. Rev*., **66,** 671–701.
2. A.A. Neyfakh et al. (1991) Efflux-mediated multidrug resistance in *Bacillus subtilis*: similarities and dissimilarities with the mammalian system, *Proc. Natl. Acad. Sci. USA,* **88,** 4781–4785.
3. M. Ahmed et al. (1995) Two highly similar multidrug transporters of Bacillus subtilis whose expression is differentially regulated. *J. Bacteriol,* **177**, 3904–3910.
4. M. Ahmed et al. (1994) A protein that activates expression of a multidrug efflux transporter upon binding the transporter substrates. *J. Biol. Chem.* **269**, 28506–28513.
5. D.P. Woolridge et al. (1997) Efflux of the natural polyamine spermidine facilitated by the Bacillus subtilis multidrug transporter Blt. *J. Biol. Chem.* **272**, 8864–8866.
6. D.P. Woolridge et al. (1999) Characterization of a novel spermidine/spermine acetyltransferase, BltD, from *Bacillus subtilis*. *Biochem J*. **340**, 753–758.

Molecular Evolution of a Complex Signal Transduction System

Kristin Wuichet, Igor Zhulin

Joint Institute for Computational Sciences, Oak Ridge National Laboratory – University of Tennessee, Oak Ridge TN 37831 USA, ijouline@utk.edu

The molecular machinery that controls chemotaxis in bacteria is substantially more complex than any other signal transduction system in prokaryotes, and its origins and variability among living species are unknown. We found that this multi-protein "chemotaxis system" is present in most prokaryotic species and evolved from simpler two-component regulatory systems that control prokaryotic transcription (1). We discovered, through genomic analysis, signaling systems intermediate between two-component systems and chemotaxis systems. Evolutionary genomics established central and auxiliary components of the chemotaxis system. While tracing its evolutionary history, we also developed a classification scheme that reveals more than a dozen distinct classes of chemotaxis systems, enabling future predictive modeling of chemotactic behavior in unstudied species.

1. K. Wuichet & I.B. Zhulin (2010) Origins and diversification of a complex signal transduction system in prokaryotes, Science Signaling, 3: ra50.

Deciphering mechanisms of miRNA action on translation by mathematical modeling

Andrei Zinovyev[1], Nadya Morozova[2], Emmanuel Barillot[1], Annick Harel-Bellan[2], Alexander Gorban[3]

[1]*Institut Curie, France, andrei.zinovyev@curie.fr*
[2]*CNRS FRE 2944, France*
[3]*University of Leicester, United Kingdom*

Protein translation is a multistep process which can be modeled as a cascade of biochemical reactions (initiation, ribosome assembly, elongation, etc.), the rate of which can be regulated by small non-coding microRNAs through multiple mechanisms. In the current literature, there are nine different mechanisms of miRNA-mediated translation repression described. However, it remains unclear what mechanisms of microRNA action are the most dominant: moreover, many experimental reports deliver controversal messages on what is the concrete mechanism actually observed in the experiment.

In this paper we suggest a mathematical approach helping to re-classify known biological mechanisms based on the types of dominant asymptotic solutions of a mathematical model of miRNA-mediated translation repression and the associated qualitatively different types of dynamical behavior of measurable quantities. This classification aims at providing a protocol for discriminating between multiple possible mechanisms of miRNA action on protein translation.

A mathematical model describing the effect of all nine miRNA-mediated mechanisms of translation repression was constructed. It was found that the model has 7 types of asymptotic solutions in the limit of well-separated constants. These types can be mapped onto known biological miRNA action mechanisms but not one to one. We suggest that this classification reflects better the biochemical nature of miRNA mechanisms, since those biological mechanisms corresponding to the same dynamical type might be indistinguishable from the point of view of measurable dynamical variables (total amount of mRNA, total amount of protein, mean number of ribosomes sitting on a translated mRNA).

Our analysis of the transient protein translation dynamics shows that it gives enough information to verify or reject a hypothesis about a particular molecular mechanism of microRNA action on protein translation. In addition, we speculate on evolutionary advantages of co-existence and co-operation of several alternative mechanisms of miRNA action even for the same miRNA-target pair.