International Conference
SocBiN Bioinformatics 2016

June 14–16, 2016

Moscow, 2016

With support from
Russian Science Foundation,
grant 15-14-3002

We are very grateful for the support in the organization of the Conference to

ISBN 978-5-4439-0110-7

Big data analysis in bioinformatics and applications
June 11–13, Moscow

http://bioinformatics2016.fbras.ru/conference/school

PROGRAM

Working languages are Russian and English

11.06.16

Venue: Institute of Bioengineering RAS (60-letiya Oktyabrya, 7/1)
Basics of NGS data analysis

8.30 Registration
8.50 Welcome remarks
9.00 Artem Kasianov. Data preprocessing and genome assembly (Russian)
10.30 Pavel Mazin. Computational analysis of RNA-Seq data (Russian)
12.00 *Coffee break*
12.20 Ivan Kulakovskiy. Digging through high-throughput data on protein-DNA interactions with sequence motif analysis (Russian)
13.50 *Lunch*
15.00 Alexander Favorov. Genometricorr: spatial correlation of genome-wide interval datasets (Russian)
15.45 *Coffee break*
16.05 Alexey Sokolov. Ancient DNA sequencing (Russian)

12.06.16

Venue: Institute of Bioengineering RAS (60-letiya Oktyabrya, 7/1)
Biological data analysis

10.00 Yulia Medvedeva. Computational epigenomics (Russian)
11.30 *Coffee break*
11.50 Ivan Antonov. Ab initio prediction of programmed ribosomal frameshifting in prokaryotes (Russian)
13.20 *Lunch*
14.30 Artem Artemov. Dna methylation and bisulfite sequencing data analysis (Russian)
16.00 *Coffee break*
16.20 Ekaterina Khrameeva. Chromatin structure (Russian)

13.06.16

Venue: Kluch (Rochdelskaya, 15)

Bioinformatics and practical applications

9.00 Carsten Daub. Going from measuring gene expression to understanding gene regulation and its relevance for medical problems (English)

10.30 Valentina Boeva. Analysis of epigenetics and chromatin states in normal and cancer cells (English)

12.00 *Coffee break*

12.20 Finn Drablos. Statistical analysis of genomic tracks (English)

13.50 *Lunch break*

15.00 Johannes Soeding and Martin Steinegger. MMseqs software suite for fast and deep clustering and searching of large protein sequence sets (English)

16.30 Vsevolod Makeev. Mathematical objects in data analysis in bioinformatics (English)

18.00 *Coffee break*

18.20 Arne Elofsson. Structure annotation of proteomes (English)

19.50 Concluding remarks

SocBiN Bioinformatics 2016
June 14–16, Moscow
http://bioinformatics2016.fbras.ru/

PROGRAM

14.06.16

9.00 Registration
9.20 Welcome remarks

9.30 Section 1: Cancer genetics and epigenetics
Chair: Yulia Medvedeva

9.30 Valentina Boeva (Invited speaker). Computational analysis of epigenetic landscape and chromatin states in cancer: Deciphering the rewiring rules of transcriptional regulatory networks in neuroblastoma
10.15 Ilya Vorontsov. Selection of somatic mutations within transcription factor binding motifs in human cancers
10.40 Tatyana Gorbacheva. Free circular DNA as a biomarker of metastasis

11.05 *Coffee break*

11.30 Section 2: From sequencing to gene regulation I
Chair: Carsten Daub

11.30 Albin Sandelin (Invited speaker). The promoter and enhancer landscape of inflammatory bowel disease
12.15 Finn Drablos (Invited speaker). How to prepare genes for rapid activation

13.00 *Lunch break*

14.30 Section 3: From sequencing to gene regulation II
Chair: Albin Sandelin

14.30 Carsten Daub (Invited speaker). Gene regulation underlying insulin response and its dys-regulation in obesity
15.15 Ekaterina Khrameeva. Dictyostelium discoideum chromosomes are partitioned into domains functionally similar to topologically associating domains of higher eukaryotes
15.40 Yulia Medvedeva. Chromatin changes induced by DNA demethylation

16.05 *Coffee break*

16.30 Section 4: RNA bionformatics and regulation
Chair: Vsevolod Makeev

16.30 Ivan Antonov. Predicting antisense interactions of long noncoding RNAs in human cells
16.55 Anna Obraztsova. Novel comparative genomic approach for detecting non-homologous RNA regulatory elements
17.30 Poster section and welcome reception

15.06.16

10.00 Section 5: Molecular evolution
Chair: Johannes Soeding

10.00 Elena Osipova. Evolution history of lipoxygenase pathway enzymes
10.25 Irina Medvedeva. How are protein functional sites encoded by exon structure in Metazoa
10.50 Maria A. Andrianova. Human mismatch repair system corrects errors produced during lagging strand replication more effectively
11.15 Alexander Panchin. Methylation and preservation of CpG dinucleotides in human CpG islands

11.40 *Coffee break*

12.00 Section 6: Structural bioinformatics
Chair: Arne Elofsson

12.00 Arne Elofsson (Invited speaker). PconsC3: Improved contact predictions for smaller families
12.45 Alexander Osypov. Electrostatics as a new old factor of the natural selection in genome
13.10 Yaroslav Popov. StructAlign — a program for alignment of structures of DNA-protein complexes
13.35 Olga Zanegina. Search of conserved features in protein-DNA complexes via Nucleic acid — Protein Interaction DataBase (NPIDB)

14.00 *Lunch break*

15.00 Section 7: Motifs in the sequences
Chair: Finn Drablos

15.00 Johannes Soeding (Invited speaker). Bayesian higher-order models consistently outperform PWMs at predicting regulatory motifs in nucleotide sequences
15.25 Ivan Kulakovskiy. HOCOMOCO collection of transcription factor binding sites models: Expansion, enhancement and practical applications
15.50 Nurbubu Moldogazieva. Short linear motifs derived from fetoplacental proteins: Bioinformatics and molecular dynamics simulation study

16.15 *Coffee break*

16.35 Section 8: Translation and regulation
Chair: Ivan Kulakovskiy

16.35 Pavel Baranov (Invited speaker). Through the ribosome profiling, darkly: Can mRNA remember and ribosomes foresee?
17.50 Vassily Lyubetsky. Ribosome reinitiation at leader peptides increases translation of bacterial proteins
18.15 Irina Eliseeva. Can transcription determine future of mRNAs? A case study on mTOR translational control in mammals.

19.30 *Conference dinner*

16.06.16

10.00 Section 9: Bioinformatics in 3D
Chair: Pavel Baranov

10.00 Nadia Chuzhanova (Invited speaker). Mutability in 3D
10.45 Sofia Mariasina. Bioinformatics approaches in NMR structure determination of methyltransferase WBSCR27

11.10 *Coffee break*

11.30 Section 10: Populational genetics
Chair: Ivan Antonov
11.30 Irina Tsvetkova. Genetic structure of *Streptococcus pneumoniae* population
11.55 Tatyana Gromovykh. The problem of the genetic stability of the strain *Gluconacetobacter hansenii* GH 1-2008
12.25 Concluding remarks

12.40 *Lunch*
14.00 *Moscow excursion*

POSTERS

Anna Petukhova. Bioinformatic approaches to building of gene regulatory networks

Anna Lioznova. Regulatory role of single CpG methylation

Praharshit Sharma. FastQ-ome: A random forest ensemble of FastQ reads as decision trees

Denis Vorobyev. Two methods to calculate P-value of RNA of a definite shape

Artem Kasianov. Identifying genes of antimicrobial peptides in transcriptomes of *Triticum kiharae* Dorof. et Migush

Artem Artemov. Differential DNA methylation of ion pump genes between marine and freshwater sticklebacks

Alexey Sokolov. DNA study of the early Bronze Age humans in the North Caucasus reveals their dual connection with Near East and Central Europe

Anatoly Bragin. Computer analysis of RNA-seq data computer of laboratory rats with aggressive behavior

Tatyana Savelieva. Transcription factors regulating gene expression in different cell types of moss *Physcomitrella Patens*

SCHOOL
Big data analysis in bioinformatics and applications

Ab initio prediction of programmed ribosomal frameshifting in prokaryotes

Ivan Antonov

Institute of Biotechnology, Research Centre of Bioengeneering, RAS, Moscow, Russia

Programmed ribosomal frameshifting (PRF) is a "recoding" event during translation when a ribosome changes the initial reading frame at a specific location in mRNA. PRFs have been observed in all domains of life and in many cases they serve a biologically important function. I would like to discuss an approach that allowed us to identify new PRFs in bacteria. This work was published in 2013 (PMID:23649834). In short, we used an ab initio frameshift prediction program GeneTack to screen 1,106 complete prokaryotic genomes identifying 206,991 genes with frameshifts (fs-genes). Our final goal was to determine if a frameshift transition was due to (i) a sequencing error, (ii) an indel mutation or (iii) a recoding event. We grouped 102,731 fs-genes into 19,430 clusters based on sequence similarity between protein products, conservation of position of predicted frameshift, and its direction. In 146 clusters with total of 4,730 fs-genes we detected conserved motifs located near frameshifts characteristic for programmed frameshifts. To test the predictions, experiments were performed with cassettes of predicted frameshift-producing sequences of fs-genes residing in 20 clusters. PRF with higher than 10% efficiency was observed for four clusters.

DNA methylation and bisulfite sequencing data analysis

Artem Artemov

Lomonosov Moscow State University, Moscow, Russia
Institute of Biotechnology, Research Centre of Bioengeneering RAS, Moscow, Russia

DNA methylation and hydroxymethylation represent an important level of epigenetic regulation. I will explain how DNA methylation is different from other epigenetic modifications and why it deserves to be studied in detail. We will discuss some important recent studies highlighting the role of DNA methylation and hydroxymethylation in cancer, in epigenetic inheritance, in imprinting and as the best marker of aging. I will briefly introduce available methods of DNA methylation profiling including BS-seq, RRBS-seq, meDIP, methylation arrays, restriction-based methods and we will focus on analysis of bisulfite sequencing data.

Analysis of epigenetics and chromatin states in normal and cancer cells

Valentina Boeva
INSERM, Institut Cochin, France

I will talk about how high throughput sequencing techniques help annotate different chromatin states (active and poised promoters, enhancers, transcribed gene bodies etc), how these states can be reorganized in cancer. I will also talk about several bioinformatics techniques for the analysis of chromatin profiles in cancer and normal cells.

Going from measuring gene expression to understanding gene regulation and its relevance for medical problems

Carsten Daub
Karolinska Institutet, Sweden

Analysis of omics data from complex medical datasets is challenging. The lecture will exemplify and discuss how specific findings with biological significance can be obtained from such complex data and complex experimental design. We will further discuss criteria for a good experimental design for complex omics data.

Statistical analysis of genomic tracks

Finn Drabløs
Department of Cancer Research and Molecular Medicine, NTNU, Norwegian University of Science and Technology, Trondheim, Norway

The presentation will show how we can do statistical analysis of genomic tracks, with focus on the Genomic HyperBrowser <https://hyperbrowser.uio.no/hb/> (PMID: 21182759). A genomic track is a linear representation of a genome-wide property, with coordinates relative to a reference genome. This means that we easily can display a collection of tracks, representing features like histone modifications, GC content or binding of transcription factors, for visual comparison in for example the UCSC Genome Browser <https://genome.ucsc.edu/>. The Genomic HyperBrowser takes this one step further, and enables statistical comparisons of tracks. This can be used for testing for example whether binding sites for a given transcription factor overlaps with CpG islands more often than expected by random chance.

Structure annotation of proteomes

Arne Elofsson
Arne Elofsson Lab, Karolinska Institutet, Sweden

I will describe the progress and status on how to structually annotate complete proteomes. We will discuss progress in sequence alignment and contact prediction methods.

Genometricorr: spatial correlation of genome-wide interval datasets

Alexander Favorov

Department of Oncology, Division of Biostatistics and Bioinformatics, Johns Hopkins University School of Medicine, Baltimore, USA
Vavilov Institute of General Genetics RAS, Moscow, Russia

Genetic data are generally arranged spatially along chromosomes, and in many genetic datasets, each item has a location in the genome, and spatial proximity often correlates with functional relationship. When comparing two genetic datasets, we can take advantage of spatial measures to test whether genetic datasets are independent of each other with regards to position on the genome, or whether they are situated in a mutually nonrandom way.

In this talk I will present our R package GenometriCorr that performs several carefully chosen spatial tests of independence on genome-wide data.

Data preprocessing and genome assembly

Artem Kasianov

Vavilov institute of general genetics RAS, Moscow, Russia

Main theme of the talk will be initial processing of sequencing data. In first part, you will learn about main file formats, sources of sequencing errors, reads quality checking and error correction methods. Second part will be devoted to genome de novo assembly. De Bruijn and overlap graphs, assembly quality checking a verification will be discussed.

Chromatin structure

Ekaterina Khrameeva

Skolkovo Institute of Science and Technology, Skolkovo, Russia

Over the last decade, development of methods based on the chromosome conformation capture technology (3C) have enabled genome-wide mapping of chromosome interactions in 3D. We will discuss general principles of chromosome spatial organization in nuclear space, methods to capture chromosome conformation (3C, 4C, 5C, HiC), and guidelines for analyzing and interpreting data obtained with genome-wide methods such as Hi-C and 3C-seq that rely on deep sequencing. Also, we will learn how to annotate topologically associating domains (TADs) in chromatin, and review several recent studies on their functional importance and mechanisms of formation.

Digging through high-throughput data on protein-DNA interactions with sequence motif analysis

Ivan Kulakovskiy

Vavilov institute of general genetics RAS, Moscow, Russia
Engelhardt Institute of Molecular Biology RAS, Moscow, Russia

First, we shall provide an introduction to computational sequence motif analysis in regulatory genomics focusing on prediction of transcription factor binding sites in genomic regulatory regions. Next, we shall highlight the synergy between motif analysis and modern high-throughput data and discuss whether careful dry-lab analysis is necessary for proper interpretation of data from chromatin immunoprecipitation (ChIP-Seq) experiments. Finally, selected case studies will be presented.

Mathematical objects in data analysis in bioinformatics

Vsevolod Makeev

Vavilov Institute of General Genetics RAS, Moscow, Russia
Engelhardt Institute of Molecular Biology RAS, Moscow, Russia
Moscow Institute of Physics and Technology, Dolgoprudny, Russia

Computational analysis of RNA-Seq data

Pavel Mazin

Skolkovo Institute of Science and Technology, Skolkovo, Russia

I will talk about RNA-seq data processing: mapping, quality control, transcript assembly, discovery of differentially expressed and alternatively spliced genes. Also, I will discuss visualization and functional analysis of RNA-seq data. I will introduce students to widely used software and packages for RNA-seq data analysis: tophat2/hisat2, cufflinks, htseq-count, interProScan, edgeR, goseq/GOstat.

Computational epigenomics

Yulia Medvedeva

Institute of Biotechnology, Research Centre of Bioengeneering, RAS, Moscow, Russia
Vavilov Institute of General Genetics RAS, Moscow, Russia

During my presentation we will discuss epigenetic processes and their role in regulation of gene expression. Then we will go through various NGS-based methods to determine epigenetic profiles of the cell. In particular we will cover specifics of ChIP-seq data processing for histone modifications. Further processing, including chromatin domain detection will also be covered. We also will discuss methods and reasons to search for the open chromatin regions. A lot of attention will be given to "best practices" at the levels of both experiment planning and data processing.

MMseqs software suite for fast and deep clustering and searching of large protein sequence sets

Johannes Soeding and Martin Steinegger
Max-Planck-Institute for Biophysical Chemistry, Goettingen, Germany

Motivation: Sequence databases are growing fast, challenging existing analysis pipelines. Reducing the redundancy of sequence databases by similarity clustering improves speed and sensitivity of iterative searches. But existing tools cannot efficiently cluster databases of the size of UniProt to 50% maximum pairwise sequence identity or below. Furthermore, in metagenomics experiments typically large fractions of reads cannot be matched to any known sequence anymore because searching with sensitive but relatively slow tools (e.g. BLAST or HMMER3) through comprehensive databases such as UniProt is becoming too costly. Results: MMseqs (Many-against-Many sequence searching) is a software suite for fast and deep clustering and searching of large datasets, such as UniProt, or 6-frame translated metagenomics sequencing reads. MMseqs contains three core modules: a fast and sensitive prefiltering module that sums up the scores of similar *k*-mers between query and target sequences, an SSE2- and multi-core-parallelized local alignment module, and a clustering module. In our homology detection benchmarks, MMseqs is much more sensitive and 4 to 30 times faster than UBLAST and RAPsearch, respectively, although it does not reach BLAST sensitivity yet. Using its cascaded clustering workflow, MMseqs can cluster large databases down to ~30% sequence identity at hundreds of times the speed of BLASTclust and much deeper than CD-HIT and USEARCH. MMseqs can also update a database clustering in linear instead of quadratic time. Its much improved sensitivity-speed trade-off should make MMseqs attractive for a wide range of large-scale sequence analysis tasks.

Ancient DNA sequencing

Alexey Sokolov
Institute of Biotechnology, Research Centre of Bioengeneering, RAS, Moscow, Russia

Typical ancient DNA workflow includes several steps. On the first step it is required to map reads on target genome and perform basic control of quality. For this step it is required to use PALEOMIX pipeline. The PALEOMIX pipeline is a set of pipelines and tools designed to aid the rapid processing of High-Throughput Sequencing (HTS) data, starting from de-multiplexed reads from one or more samples, through sequence processing and alignment, followed by genotyping and phylogenetic inference on the samples. Also on this stage it is necessarily to access contamination level with contamMix. On the second stage it is required to do snp calling analysis, for this step one can use any available SNP caller. And on the final step one can use HaploGrep in order to find haplogroups of target samples. Also on the final step it might be appropriate to exclude modern reads from analysis using using PmdTools in order to eliminate conflicting SNP's.

Human mismatch repair system corrects errors produced during lagging strand replication more effectively

Maria A. Andrianova*[1,2,3], Georgii A. Bazykin[1,2,3,4], Sergey I. Nikolaev[5], and Vladimir B. Seplyarskiy[1,3]

* baranova.mariia@gmail.com
1 Institute for Information Transmission Problems of the Russian Academy of Sciences (Kharkevich Institute), Bolshoi Karetny pereulok 19, Moscow 127994, Russia
2 Moscow State University, Leninskie gory 1, Moscow 119234, Russia
3 Pirogov Russian National Research Medical University, Ostrovitianov str. 1, Moscow 117997, Russia
4 Skolkovo Institute of Science and Technology, Skolkovo 143026, Russia
5 Department of Genetic Medicine and Development, University of Geneva Medical School, 1 Rue Michel-Servet, 1211 Geneva, Switzerland; Institute of Genetics and Genomics in Geneva, 1211 Geneva, Switzerland

Mismatch repair (MMR) is one of the main systems maintaining fidelity of replication. Different effectiveness in correction of errors produced during replication of the leading and the lagging DNA strands was reported in yeast, but this effect is poorly studied in humans. Here, we use MMR-deficient (MSI) and MMR-proficient (MSS) cancer samples to investigate properties of the human MMR. MSI, but not MSS, cancers demonstrate unequal mutation rates between the leading and the lagging strands. The direction of strand asymmetry in MSI cancers matches that observed in cancers with mutated exonuclease domain of polymerase δ, indicating that polymerase δ contributes more mutations than its leading-strand counterpart, polymerase ε. As polymerase δ primarily synthesizes DNA during the lagging strand replication, this implies that mutations produced in MMR efficient cells during lagging strand replication are repaired by the MMR more effectively, compared to those produced on the leading strand.

Predicting antisense interactions of long noncoding RNAs in human cells

Ivan Antonov*[1]

* ivan.antonov@gatech.edu
1 Research Center of Biotechnology RAS, Moscow, Russia

Long noncoding RNAs (lncRNAs) are a large and diverse class of transcribed RNA molecules with a length of more than 200 nucleotides that do not encode proteins. The discovery of thousands of lncRNAs in mammals raised a question about their

functionality. Due to functional diversity the role and/or molecular mechanism of only few hundred lncRNAs have been determined by the date. Particularly, it has been shown that some of them function post-transcriptionally via formation of intermolecular RNA-RNA duplexes. The primary aim of this study is to bioinformatically address novel lncRNA functions by predicting RNA-RNA interactions transcriptome-wide. To search for potential antisense partners for a given non-coding RNA, existing large-scale studies utilized sequence alignment tools (such as BLASTn) without taking into account RNA secondary structure and interaction energy, crucial for RNA-binding. To compensate for this disadvantage co-folding of two RNAs (the query lncRNA versus each of the RNAs in the transcriptome) into minimal free energy (MFE) structure using thermodynamics-based methods (e.g. bifold) can be used. Unfortunately, this task is not computationally feasible on the transcriptome-wide level. In this work we developed a new pipeline, called ASSA ("AntiSense Search Approach"), which reduces running time by fast identification of putative antisense sites by a sequence alignment tool BLASTn followed by verification of each potential interaction by bifold. In our pipeline we automated selection of the initial set of putative antisense sites (i.e. optimized thresholds for BLASTn search), estimated statistical significance (p-value) of antisense interaction energy and the length of the flanking sequences to putative site for validation by bifold. ASSA was capable of predicting 26 out of the 29 known functional RNA-RNA interactions (both cis and trans) in human and mouse transcriptomes. We also applied ASSA to publicly available data from knockdown experiments of 49 murine lncRNAs. We identified four lncRNAs with statistically significant overlap between the ASSA predictions and the differentially expressed genes observed in the experiment, suggesting possible molecular mechanism for these long noncoding RNAs. Finally, we have shown that ASSA could be used for ab initio prediction of regulatory lncRNA-RNA interactions. This option could be particularly useful for wet-lab biologists as it suggests targets for experimental studies.

This work was supported by RSF grant 15-14-30002.

Differential DNA methylation of ion pump genes between marine and freshwater sticklebacks

Artem V. Artemov*[1], Nikolai S. Mugue[2], Sergey M. Rastorguev[3], Alexander M. Mazur[1], Svetlana V. Tsygankova[3], Artem V. Nedoluzhko[3], Yulia A. Medvedeva[1,4], and Egor B. Prokhortchouk[1]

* artem.v.artemov@gmail.com
1 Institute of Bioengineering, Research Center of Biotechnology RAS, Moscow, Russia
2 Russian Federal Research Institute of Fisheries and Oceanography, Moscow, Russia
3 National Research Center 'Kurchatov Institute', Moscow, Russia
4 Vavilov Institute of General Genetics RAS, Moscow, Russia

Three-spined stickleback (*Gasterosteus aculeatus*) represents a convenient model to study microevolution — adaptation to freshwater environment. While genetic adaptations to freshwater are well-studied, epigenetic adaptations attracted little attention.

In this work, we investigated the role of DNA methylation in the adaptation of marine stickleback population to freshwater conditions. DNA methylation profiling was performed in marine and freshwater populations of sticklebacks, as well as in marine sticklebacks placed into freshwater environment and freshwater sticklebacks placed into seawater. For the first time, we demonstrated that genes encoding ion channels kcnd3, cacna1fb, gja3 are differentially methylated between marine and freshwater populations. We also showed that after placing marine stickleback into fresh water, its DNA methylation profile partially converges to the one of a freshwater stickleback. This suggests that immediate epigenetic response to freshwater conditions can be maintained in freshwater population. Interestingly, we observed enhanced epigenetic plasticity in freshwater sticklebacks that may serve as a compensatory regulatory mechanism for the lack of genetic variation in the freshwater population. Some of the regions that were reported previously to be under selection in freshwater populations also show differential methylation. Thus, epigenetic changes might represent a parallel mechanism of adaptation along with genetic selection in freshwater environment.

Through the ribosome profiling, darkly: Can mRNA remember and ribosomes foresee?

Pavel Baranov*[1]

* p.baranov@ucc.ie
1 University College Cork, Ireland

Ribosome profiling (Ribo-Seq) technique has been developed seven years ago to assess gene expression at the transcription and translation level at the scale of the entire cell transcriptome. The ability of the technique to produce unprecedentedly detailed quantitative characterization of gene expression on the global scale made it popular. The applications of the technique prompted researchers to reconsider the current views on how protein coding information is organized in the genomes and on how protein synthesis is carried out and regulated in the cells. However, ribosome profiling data are highly heterogeneous and difficult to analyse due to the presence of sporadic technical noise. We recently developed a simple smoothing technique (RUST for Ribo-Seq Unit Step Transformation) that allows to circumvent these problems. We showed that RUST can be used for assessing the quality of datasets and estimating how properties of mRNA effect the speed of elongating ribosomes. The analysis of human ribo-seq data revealed an unusual feature of a specific mRNA translation that prompted us to discover a molecular mechanism of memory formation in translated mRNA molecules. The analysis of ribo-seq data obtained in ciliates Euplotes revealed several thousands of highly efficient productive ribosomal frameshifting events. Surprisingly, frameshifting in Euplotes does not require apparent sequence signals typical for programmed ribosomal frameshifting in other organisms.

Computational analysis of epigenetic landscape in cancer: Deciphering the rewiring rules of transcriptional regulatory networks in neuroblastoma

Valentina Boeva*[1]

* valentina.boeva@inserm.fr
1 INSERM, Institut Cochin, France

Epigenetic landscape, i.e., positioning of diverse epigenetic marks along the genomic DNA, is known to evolve during tumor development and progression. For instance, as a consequence of specific genetic events (mutations or structural variants) or due to the general rewiring of the transcriptional regulatory networks, oncogenes often gain de novo active regulatory elements (enhancers and super-enhancers) in cancer cells. Overexpression of certain components of repressive complexes (e.g., the EZH2 component of the Polycomb Repressive Complex 2) or mutations in genes coding for chromatin remodeling proteins (e.g., MLL2 and MLL3) can also result is epigenetic silencing of tumor suppressor genes or epigenetic activation of oncogenes. Thus, analysis of changes in the epigenetic landscape using the ChIP-seq technique can show extremely useful to get insights into the mechanisms of cancer development and progression. In addition, information about the epigenetic profiles obtained with ChIP-seq can in some cases provide prognostic markers for patient stratification and, for sub-groups of patients, suggest efficient treatment affecting epigenetic states. For instance, a wide panel of 'epigenetic' compounds (inhibitors of EZH2, HDAC, BRD4 and CDK7) has been shown to be efficient in a number of cancer cell lines and mouse models. In my presentation, I will show how ChIP-seq data can be used in cancer studies to annotate chromatin states and assess changes of the epigenetic landscape. In particular, I will present methods we have developed to normalize profiles and detect regions of enrichment in chromatin marks for ChIP-seq datasets coming from cancer samples and also identify regions of differential enrichment. Our methods account for copy number aberrations of cancer genomes, different technical biases and variable signal-to-noise ratio between different experiments. I will also present our unpublished results on the analysis of epigenetic profiles of 20 neuroblastoma cell lines, where we characterized the most potent de novo super-enhancers and associated them with neuroblastoma development.

Computer analysis of RNA-seq data computer of laboratory rats with aggressive behavior

Anatoly O. Bragin*[1], Vladimir N. Babenko[1], Anastasia M. Spitsina[1], Ekaterina V. Kulakova[1], Irina V. Medvedeva[1], Irina V. Chadaeva[1], Yuriy L. Orlov[1], and Arcady L. Markel[1]

* ibragim@bionet.nsc.ru

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

Aggressive behavior in our society is one of the most urgent social problems. It is concluded from the studies conducted in recent decades that aggression should be considered as a special feature of human behavior based also on the genetic and sociobiological history of the development of humans as a biologic species, of an individual, and of social group and environment. Therefore, the investigation of the molecular mechanisms determining the aggressive behavior is an important task. With respect to the notion that aggression is a general biological phenomenon stemming from deep evolutionary roots, the investigation of aggressive behavior demands that not only the aggressive behavior of a human be analyzed but also the corresponding features of animal behavior. For analyzing of genetics factor of aggression we have use aggressive and tame rat line. These rat lines have selected in Institute of cytology and genetics SB RAS. We have use three rat brain area: hypothalamus, ventral tegmental area and midbrain raphe nuclei. RNA sequencing of the tissues was carried out. This was followed by mapping on rat genome and differently expressed genes prediction by TopHat and Cufflinks programs. We use DAVID tools to find biological processes with rat differentially expressed genes is overrepresented. We have use ANDvisio tools for associative gene network reconstruction. It was found that differentially expressed genes of the rats were associated with behavior. This program use information of genes regulation from different databases and use textmining technology. Using computer analysis of the three brain tissues of aggressive and tame rats the differentially expressed genes were found. The differentially expressed genes are involved in biological processes responsible for neuronal signal transduction, behavior, neuron development and other processes of neuronal activity. The association of genes Gabrd, Shank3, Egr1 — involved in human mental diseases — with aggression was confirmed. Using ANDvisio system the associative genes networks of the differentially expressed genes with molecular interaction were reconstructed.

Mutability in 3D

Nadia Chuzhanova*[1]

* nadia.chuzhanova@ntu.ac.uk

1 School of Science and Technology, Nottingham Trent University, Clifton Lane, Nottingham, NG11 8NS, UK

Advances in techniques for capturing three-dimensional chromosome conformations (3C) have been prompted the view that direct long-range interactions occur between

gene promoters and distal genomic regions, bringing them into close spatial proximity through looping interactions, thereby explaining the impact of pathological mutations that are known to occur at some considerable distance from the genes whose function they influence. In this study we hypothesised that both recurrent and non-recurrent non-coding single nucleotide variants exert their influence on a target gene(s) either directly, via long-range looping interactions between fragments that harbour alterations and a target (often unknown) gene promoter(s), or by 'propagating' via the network of interactions governed by the 3-dimensional architecture of the human genome.

In silico approach, based on this hypothesis, was successfully employed in the context of the prediction of potential remotely-acting regulatory elements for several inherited diseases. Remotely-acting regulatory elements for the *NF1* gene were predicted and sequenced in 47 patients with neurofibromatosis type 1, lacking mutations in either *NF1* or *SPRED* genes. Three sequence variants were found in the predicted regions in 5/47 patients with NF1. Remotely-acting regulatory elements for the *SMCHD1* gene were predicted and sequenced in 229 patients with facioscapulohumeral dystrophy FSHD, lacking mutations in the *SMCHD1* gene. Three sequence variants were found in the predicted region in 184/229 patients with FSHD.

This approach demonstrates a novel means to screen for disease-relevant mutations that reside beyond the immediate vicinity of a given disease gene. It therefore promises not only to be useful in investigating disorders in which mutations may occur in remotely-acting regulatory elements but also in identifying the causative non-coding mutations found by GWAS that are often distant from their target genes.

Gene regulation underlying insulin response and its dys-regulation in obesity

Carsten Daub*[1]

* carsten.daub@ki.se
1 Karolinska Institutet, Sweden

Human white adipose tissue responds to insulin by taking up blood glucose. In obese individuals, this glucose uptake is impaired and can lead to type 2 diabetes. However, up to 30% of obese individuals display normal glucose levels and respond well to insulin. We performed Cap Analysis Gene Expression (CAGE) transcriptome profiling of obese insulin resistant and insulin sensitive individuals as well as in lean controls both at fasting level and at high blood sugar level (hyperglycemia). We identified the insulin response mediating genes as well as they key regulation events driving these responses and their dys-regulation in obese individuals including transcription factors and enhancers.

How to prepare genes for rapid activation

Finn Drabløs*[1]

* finn.drablos@ntnu.no
1 Department of Cancer Research and Molecular Medicine, NTNU, Norwegian University of Science and Technology, Trondheim, Norway

Some genes show a very rapid activation of gene expression after stimulation, and this is known as the immediate-early response. The maximum expression level may be reached after just minutes and without the need for protein synthesis of for example transcription factors. It is therefore interesting to have a good understanding of how such genes are activated. It has previously been shown that immediate-early response genes often are in an epigenetic bivalent or poised state, where both active and repressive epigenetic signatures are present at the same time; such genes are initially repressed, but may rapidly be activated. We have generated a well-curated consensus set of genes showing rapid activation of gene expression after different types of stimulation. The genomic locations of the identified genes have been matched against genomic features known to be important for gene regulation, such as binding of transcription factors, insulators (CTCF) and cohesin, epigenetic state of chromatin and DNA methylation. The analysis also includes ChIA-PET data on chromatin interactions to analyse gene regulation through DNA looping. Our analysis highlights the importance of interaction between promoters of immediate-early response genes and strong distal enhancers for gene activation, and how in particular cohesin may facilitate the process.

Can transcription determine future of mRNAs? A case study on mTOR translational control in mammals

I. A. Eliseeva*[1], I. E. Vorontsov[2,3], and I. V. Kulakovskiy[2,3]

* yeliseeva@vega.protres.ru
1 Institute of Protein Research RAS, Pushchino, Russia
2 Engelhardt Institute of Molecular Biology RAS, Moscow, Russia
3 Vavilov Institute of General Genetics RAS, Moscow, Russia

Transcriptional regulation of gene expression can determine mRNA stability and localization in yeast. It is an open question whether there are similar machinery in higher eukaryotes, and, in particular, whether translational efficiency of particular mRNAs can be directly influenced by the transcriptional stage. In higher eukaryotes, the translation of many ribosomal and translational factors genes is controlled by the mTOR pathway that is directly involved in cell proliferation, aging, and oncogenesis. The 5' terminal oligopyrimidine sequence motif (TOP) is the specific feature of many mTOR translational targets. However, fuzziness of transcription initiation cannot always guarantee the exact location of TOP at mRNA 5' termini, and there are many mTOR targets completely lacking TOP. Other mRNAs features that are specific for

mTOR translational targets remain unclear. It is tempting to apply sequence analysis methods to identify specific transcriptional regulators that may leave imprints on transcribed mRNAs and thus determine forthcoming translational control. We utilized public CAGE and Ribo-Seq data to identify mTOR targets in human and mouse and performed sequence motif analysis of the respective promoter regions. We found binding sites of several transcription factors significantly enriched in promoters of the mTOR targets, with particular transcription factors known to possess direct RNA-binding activity or interact with other RNA-binding proteins. This suggests a principal role of transcription in mTOR translational control in higher eukaryotes.

This study was supported by Programs on "Molecular and Cell Biology" and RFBR grant 14-04-01838

PconsC3: Improved contact predictions for smaller families

Arne Elofsson*[1]

* arne@bioinfo.se
1 Arne Elofsson Lab, Karolinska Institutet, Sweden

The ultimate goal of structural bioinformatics research is to provide a complete structural map of all macromolecules and their interactions within a cell. Knowledge about the structure of proteins and other macromolecules is essential for our understanding of biological processes. Proteins do most of the work in a cell and therefore the studies of proteins structure has been an essential part of life science research during the last decades. The rapid growth of determined protein structures has made it possible to build homology models for many proteins. However, surprisingly the exponential reduction in sequencing costs has also been fundamental for the progress since it allows more distant homologies to be detected and nowadays be used to predict contacts in proteins reliably. Here, I will describe our work on the development of our contact predictor, PconsC. By using a combination of direct coupling analysis, classical machine learning and deep learning approaches we are now able to accurately predict the contacts of protein families with as little as 100 effective sequences. Further, most proteins do not act alone but through interactions with other proteins (and other molecules). Therefore, it is essential to understand not only the structure of a protein but also its interactions. Here, systems biology approaches are often used to understand what interactions are made but these studies mostly ignore the atomistic details about the interactions, i.e. how, the interactions are made. Also here the accurate prediction of inter-residue contacts can provide valuable information. I will discuss the outlook for how contact prediction can be used here.

Trends in bioinformatics and it industry

Dmitry Geraskin*[1]

* dgvvp@mail.ru
1 Management company VVP

Free circular DNA as a biomarker of metastasis

Tatyana Gorbacheva*[1]

* gorb.tanya@gmail.com
1 Voronezh State University, Russia

The development of novel techniques for the evaluation of blood biomarkers (liquid biopsies) in cancer such as circulating tumor cells (CTCs) and circulating tumor DNA (ctDNA) or tumor-specific cell free DNA (cfDNA) have changed the translational medicine as a non-invasive biomarker. Circulating biomarkers might represent both primary tumor and metastatic deposits and provide ways for investigating metastatic processes. The present study investigated the levels of patient-specific mutations in circulating cell-free DNA (cfDNA) in plasma from patients with clear-cell renal cell carcinoma. Pairs of normal-tumor tissue were sequenced, analyzed to establish the mutation profile for each patient and allele-specific PCR were conducted to monitor mutations in cfDNA in plasma. It was observed that there is a tendency of increasing the level of mutant alleles in plasma in patients with metastasis.

The problem of the genetic stability of the strain *Gluconacetobacter hansenii* GH 1-2008

T. Gromovykh*[1], I. Petrichin[1], and A. Demchenko[1]

* gromovykhtatyana@mal.ru
1 First Moscow State Medical University IM Sechenov, Russia

Bacterial cellulose (BC) has recently been studied in biomaterials fields such as a cartilage scaffold, DNA separation medium, dental implants, nerve regeneration and vascular grafts, artificial skin. The intrinsic properties of BC make it an attractive novel biomaterial. Bacterial cellulose is a polysaccharide produced by several microorganisms, particularly *Gluconacetobacter xylinus* and *Gluconacetobacter hansenii*. However it is known that the bacteria producing the bacterial cellulose (Cel + cells) genetically unstable and form mutant cells do not synthesize cellulose (Cel-mutants). The productivity of bacterial cellulose synthesis depends on the strain and the emergence of genetics Cel-mutants cells. Gene encoding the cellulose biosynthesis is organized in the form of an operon (acs). The nucleotide sequence of the region shows the presence of two open reading frames, ORF1 and ORF2. It was shown that the mutant cells (Cell-) have an insertion sequence element (IS1031A) (approximately 500 nucleotides) violating reading cellulose synthesis operon gene. Because these two cell types have distinct morphologies, they can be simultaneously quantified by assessing the number of colony forming units. The Cel- mutants are smooth colonies while the Cel+ are mucosal and rough. (*Jung Hwan Ha et al.*, 2011). Bacterial cellulose polymer properties such as crystallinity, strength, nanostructure fibers and water-holding ability dependent producer synthesizing ability. Therefore, genetic stability, associated with the presence in the population of Cel + cells is an actual problem

for quality polymer by culturing producer. Investigations were carried out with the strains *Gluconacetobacter hansenii* GH 1-2008 of the isolated us in 2008. BC was produced in plate forms in static cultures, respectively. In order to prevent the appearance of mutant cells cultivation was performed in nutrient modified media S. Hestrin and M. Schramm with the following composition, g/l: sucrose — 20, peptone — 5, yeast extract — 5, Na_2HPO_4 — 2.7, monohydrate citric acid — 1.15 and various concentrations of ethanol (0.5, 1.0, 1.5, 2.0 and 3.0%). The BC concentration was measured after 5, 10, and 15 days. The BC was harvested from the static culture by simply removing the BC plate from the culture flask and washing it thoroughly with distilled water. Microbial cells was measured after the compound was freeze-dried at −50 °C seeding on agar medium Hestrin and M. Schramm. As a result of studies found that to suppress the generation of a mutant bacterial strain *G. hansenii* GH 1-2008 optimum ethanol concentration is 2.5% or more. The production of bacterial cellulose was increased from 2.7 to 8.5 g/l, and strength. Cel + cells in the culture medium and the film is maximized in the culture period. Bacterial cellulose film, *G. hansenii* synthesized in a medium with a concentration of ethanol of less than 2.5% without ethanol and transparent, less strong and have a higher moisture content. At these concentrations of ethanol in the culture medium are detected cells Cel-mutants. Substitution in the organic nitrogen source of the nutrient medium (peptone and yeast extract) also prevents unstable Cel-mutants cells in culture medium and the film. Thus, the stability *G. hansenii* population depends on the composition of the nutrient medium containing the necessary quantity of ethanol or an organic nitrogen source.

Identifying genes of antimicrobial peptides in transcriptomes of *Triticum kiharae* Dorof. et Migush.

Artem S. Kasianov*[1], Alexey S. Kovtun, Tatyana V. Korostyleva, Ekatherine A. Istomina, Vsevolod J. Makeev, Alexander M. Kudryavtsev, and Tatyana I. Odintsova

* artem.kasianov@gmail.com

1 Vavilov institute of general genetics RAS, Moscow, Russia

Contamination of plants with pathogens is usually followed by immune response, that includes synthesis of various antimicrobial compounds and antimicrobial peptides (AMP) are the most important of them. AMPs in plants are short (< 100 a.a.), positively charged, cysteine-rich polypeptides that are able to respond to wide spectrum of pathogens. The goal of this work is identification of transcripts that code AMP in healthy and infected by Fusarium oxysporum seedlings of *Triticum kiharae* Dorof. et Migush. wheat using methods of high-throughput sequencing (NGS). The cDNA libraries were sequenced on Illumina HiSeq2000. We developed a method of in silico searching of AMPs in the obtained transcripts. Firstly, we used a method of hidden Markov models, which were created from already known AMPs from *A. thaliana* and *O. sativa*. Secondly, we wrote a program in Perl that uses regular expressions for

searching AMPs by their cysteine-rich motifs. As a result, we identified more than 500 amino acid sequences of AMPs that correspond to lipid transport proteins, defensins, hevein-like peptides and many others. The lipid transport proteins appeared to be the most widespread family in our resulting set — approximately half of the total number of AMPs. We also identified 145 transcripts of defensins and 3 transcripts coding hevein-like peptides. The majority of identified transcripts are new, previously not described in wheat. Because the genome of *T. kiharae* consists of subgenomes, which originate from *T. urartu* and *A. tauschii*, we showed, which of our identified transcripts refer to these subgenomes and which are individual for *T. kiharae*. We showed that the expression of some identified genes of AMPs enhances when wheat is infected with the pathogen. Aside from the known families of plant AMPs we identified cysteine-rich peptides with new cysteine motifs. Their role in plants needs a further investigation.

Our work is supported by RFBR grants №№ 15-04-0468015/15 and 15-29-02480.

Dictyostelium discoideum chromosomes are partitioned into domains functionally similar to topologically associating domains of higher eukaryotes

Ekaterina Khrameeva*[1]

* e.khrameeva@skoltech.ru
1 Skolkovo Institute of Science and Technology, Moscow, Russia

Recent advances enabled by the Hi-C technique have unraveled many principles of chromosomal folding that were subsequently linked to disease and gene regulation. In particular, Hi-C revealed that chromosomes of higher eukaryotes are organized into Topologically Associating Domains (TADs), evolutionary conserved compact chromatin domains that influence gene expression. To explore principles of chromosomal folding in a soil-living amoeba *Dictyostelium discoideum*, we performed Hi-C and constructed high-resolution interaction maps that revealed the presence of small (20–80 kb) loose globular domains. As in the previous studies on Drosophila and mammalian TADs, boundaries of globular domains of *D. discoideum* were enriched with actively transcribed genes and housekeeping genes, suggesting a functional similarity between globular domains of *D. discoideum* and TADs of higher eukaryotes.

HOCOMOCO collection of transcription factor binding sites models: Expansion, enhancement and practical applications

Ivan Kulakovskiy*[1]

* ivan.kulakovskiy@gmail.com
1 Engelhardt Institute of Molecular Biology, Moscow, Russia

In the latest release, HOCOMOCO COmprehensive MOdel COllection [HOCOMOCO v10, http://hocomoco.autosome.ru] provides binding models for 6 hundreds of

human and almost 4 hundreds of mouse transcription factors. The primary collection provides classic mononucleotide position weight matrices (PWMs), which were produced using ChIPMunk motif discovery tool [http://autosome.ru/ChIPMunk/]. In addition to basic mononucleotide position weight matrices (PWMs), HOCOMOCO includes dinucleotide position weight matrices based on ChIP-Seq data. For motif finding in large sequence sets we provide the command line tool, SPRY-SARUS (SuperAlphabet Representation Utilized for motif Search), able to efficiently handle both mono- and dinucleotide PWMs. ChIP-Seq data for HOCOMOCO v10 motif discovery and benchmarking was extracted from GTRD database [http://gtrd.biouml.org]. For all the models we provide quality ratings that are based on a comprehensive benchmark study or inherited from HOCOMOCO v9. We present a complete workflow used to build HOCOMOCO and discuss its practical applications and possible improvements.

This study was supported by RFBR grants 15-34-20423 and, partly, 14-04-01838.

Regulatory role of single CpG methylation

Anna V. Lioznova*[1], Abdullah Khamis[2], Artem V. Artemov[1], Vladimir B. Bajic[2], and Yulia A. Medvedeva[1,3]

* anna.lioznova@gmail.com
1 Institute of Biotechnology, Research Centre of Bioengeneering, RAS, Moscow, Russia
2 King Abdullah University of Science & Technology, Thuwal, Saudi Arabia
3 Vavilov Institute of General Genetics RAS, Moscow, Russia

The diversity of different tissues shows the ability of Nature to vary expressed genes. The regulation of gene expression has been well studied, but yet has not been understood in details. There is a number of mechanisms to control this process and DNA methylation is one of them. Typically the DNA methylation of the gene promoter suppresses the initiation of its transcription. Modern techniques allow to explore DNA methylation with single base resolution (using for example bisulfite sequencing), yet usually for the downstream analysis levels of DNA methylation are averaged along a regulatory region (e.g. promoter or CpG island).

In our study we aim to explore the relationship between gene expression and methylation of single nucleotides. We show that the methylation of particular CpGs, located in gene proximity, demonstrate significant negative correlation with the expression of the corresponding gene. We call such positions CpG traffic lights. Although some of the traffic lights are co-located with each other and demonstrate similar methylation patterns, many of them are quite different from the neighbouring CpG positions. This observation suggests that if methylation levels are averaged along the huge regions such as promoter, information about regulatory potential might be lost. Indeed, we demonstrate that CpG traffic lights are significantly overrepresented in enhancers of various types and at the exact TSS position detected by CAGE. Regions of the active histone modifications (H3K4me1, H3K4me3 and H3K27ac) are also enriched for CpG

traffic lights. We speculate that a single CpG traffic light (or a short cluster of them) is indeed a unit of transcriptional regulation by DNA methylation.

This work is partially supported by RFBR grant 14-04-00180 to YAM.

Ribosome reinitiation at leader peptides increases translation of bacterial proteins

Vassily Lyubetsky*[1,2], Semen Korolev[1], Oleg Zverkov[1], and Alexander Seliverstov[1]

* lyubetsk@iitp.ru
1 Institute for Information Transmission Problems of the Russian Academy of Sciences (Kharkevich Institute), Moscow, Russia
2 Lomonosov Moscow State University, Russia

Short leader genes usually do not encode stable proteins, although their importance in expression control of bacterial genomes is widely accepted. Such genes are often involved in the control of attenuation regulation. However, the abundance of leader genes suggests that their role in bacteria is not limited to regulation. Specifically, we hypothesize that leader genes increase the expression of protein-coding (structural) genes via ribosome reinitiation at the leader peptide in the case of a short distance between the stop codon of the leader gene and the start codon of the structural gene. For instance, in Actinobacteria, the frequency of leader genes at a distance of 10-11 bp is about 70% higher than the mean frequency within the 1 to 65 bp range; and it gradually decreases as the range grows longer. A pronounced peak of this frequency-distance relationship is also observed in Proteobacteria, Bacteroidetes, Spirochaetales, Acidobacteria, the Deinococcus-Thermus group, and Planctomycetes. In contrast, this peak falls to the distance of 15-16 bp and is not very pronounced in Firmicutes; and no such peak is observed in cyanobacteria and tenericutes. Generally, this peak is typical for many bacteria. Some leader genes located close to a structural gene probably play a regulatory role as well. Bacteria demonstrate an unusually high frequency of the 10-11 bp distance between the leader and structural genes. A less pronounced local peak is observed for the 5 bp distance. In Actinobacteria the frequency of leader genes at a distance of 10-11 bp is about 70% higher than the mean frequency within the 1 to 65 bp range; and it gradually decreases as the range grows longer. Similar peaks for the distance of 10-11 bp are observed in certain actinobacterial genera with a large number of species: Corynebacterium, Mycobacterium, and Streptomyces. Minor peaks for the 3, 5-6, 10-12, and 37 bp distances between the leader and structural genes are observed in the early diverged genus Bifidobacterium. A pronounced peak is observed in Proteobacteria as a whole as well as in individual taxonomic groups. In alpha-, beta-, epsilon-, delta-, and gamma-proteobacteria, the distance of 10-13 bp peaks; although it is less pronounced in gamma-proteobacteria; while epsilon-proteobacteria demonstrate a sharp peak for the distance of 10-11 bp. In Bacteroidetes, the distance of 11-14 bp peaks. The peaks of

leader gene frequencies at these distances are observed in Spirochaetales, Acidobacteria, the Deinococcus-Thermus group, and Planctomycetes. Notably, the peak in Planctomycetes becomes more pronounced after the genus Planctomyces is excluded. On the contrary, the peak in Firmicutes corresponds to the distance of 15-16 bp and is not very pronounced. The considered peak is absent in cyanobacteria and tenericutes, which remains true after filtering out genera with many species. An adjacent leader gene can be involved in the regulation of gene expression. Specifically, ribosome stalling at regulatory codons decreases the rate of ribosome reinitiation. In general terms, a large number of leader genes without a pronounced abundance of regulatory codons for rare amino acids assumes that most of them are not involved in regulatory mechanisms depending on the concentration of amino acids or aminoacyl-tRNA synthetases. Our data on Actinobacteria and Proteobacteria are in good agreement with the high incidence of the attenuation mechanism including its classical variant in them. Taxonomic groups Cyanobacteria, Tenericutes, and Firmicutes, which had no peak or not very pronounced peak, represent early diverged bacterial branches. In addition, there is no reliable data on the presence of attenuation mechanism in cyanobacteria and tenericutes. This also agrees with our hypothesis specified below. We propose that the leader genes in such proximity to structural genes make up a common operon with them to increase the rate of translation initiation of the structural gene product. In this case, the ribosome can start translation directly or reinitiate translation after translating the leader peptide. Thus, the leader region functions as an "antenna" on a polycistronic mRNA to increase the rate of translation initiations for the structural gene product. Different frequencies for different stop codons of leader genes can reflect the involvement of stop codons in ribosome reinitiation. In particular, the rate of ribosome release after translation can differ for stop codons since different stop codons interact with different release factors. Small quantities of leader genes at a distance of 6-9 bp can be attributed to the overlapping of this region by the purine-rich Shine-Dalgarno sequence, which interferes with the presence of any pyrimidine-containing stop codon here. This effect is valid for all stop codons. We assume no direct relationship between the distance from the leader gene to the first structural one, on the one hand, and the distance between neighboring structural genes, on the other hand. Several examples for the latter distance problem are following. There is a pronounced peak for this distance of 10-11 bp, a minor peak for the distance of 13-16 bp, and no peak for positive distances in *Escherichia coli* K-12, *Bacillus subtilis* 168 and *Synechocystis* sp. PCC 6803 (NC_000911). This pattern is similar to that revealed for leader genes. In all cases of data averaged for many species, there is a peak between neighboring structural genes for the distance of 2-4 bp, which is not observed for leader genes. The distances in the range of 6-9 bp are rare. The frequency of pairs of neighboring structural genes at the distance of 10-13 bp has a weak local maximum of considerable width. The local maximum falls on the distance of 10-11 bp in Actinobacteria and Proteobacteria; 10 bp, in Spirochaetales and Cyanobacteria; and 13 bp, in Firmicutes (which is particularly wide in this case). In Firmicutes, the position of the maximum roughly corresponds to that for leader

genes (with no account of cyanobacteria). The diagram for Proteobacteria is distant from that for *E. coli*. Thus, the peak can be pronounced in some cases or missing in other ones.

Bioinformatics approaches in NMR structure determination of methyltransferase WBSCR27

S. S. Mariasina*[1], O. A. Petrova[1], C.-F. Chang[2], T.-H. Huang[3], I. A. Osterman[1], A. B. Mantsyzov[1], P. V. Sergiev[1], and V. I. Polshakov[1]

* sm1024sm@yandex.ru
1 M. V. Lomonosov Moscow State University, Russia
2 Genomics Research Center, Academia Sinica, Taiwan
3 Institute of Biomedical Sciences, Academia Sinica, Taiwan

Protein WBSCR27 was recently discovered in human proteome. Gene which codes this protein is a part of deletion related to severe genetic disorder — Williams syndrome.

Analysis of the amino acid sequence using pBLAST showed that the closest homologue of WBSCR27 is methyltransferase WBSCR22. Similar search carried out within the RCSB Protein Data Bank showed no close homolog of the WBSCR27 among the proteins with known three dimensional structures. Knowledge of 3D structure is essential for understanding the physiological role of this protein. We decided to obtain the necessary structural information using the NMR spectroscopy techniques.

The first step in NMR data analysis is resonance assignment, i.e. determination of chemical shifts (CS) of atoms. For large biomolecules, such as proteins, this is time-consuming procedure that includes collection of series of multidimensional spectra, their processing and correlation analysis.

When substantial part of protein backbone atoms were assigned and values of their chemical shifts (CS) were determined, we performed bioinformatic analysis, facilitating determination of the remaining CS data, and allowing us to obtain information about protein secondary structure and dynamics. Secondary structure of WBSCR27 was determined using two different methods, based on the analysis of CS data and amino-acid sequence. These approaches also allowed us to obtain information about protein backbone flexibility.

On a second stage, we have used CS information to search structural analogues of WBSCR27 protein using the method SimShiftDB. Such analysis revealed similarities of WBSCR27 topology with structure of several methyltransferases, confirming the initially predicted primary function of the protein.

Thus described bioinformatic approaches facilitated us to confirm function of novel protein, to determine its secondary structure and backbone dynamics. This data will be used in subsequent structure calculation of methyltransferase WBSCR27 in solution.

How are protein functional sites encoded by exon structure in Metazoa?

Irina Medvedeva*[1]

* brukaro@gmail.com
1 Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia

Study of the relationship between the structural and functional organization of proteins and their coding genes is necessary for an understanding of the evolution of molecular systems. Functional sites represented by several amino acid residues and remoted in 3D structure are crucial for protein functionality. As the result protein functional sites are highly conserved. So analysis of the general properties of the structural organization of the functional sites at the protein level and, at the level of exon-intron structure of the coding gene is still an actual problem. To analyze functional sites encoding features amino acid residue positions were projected to the gene structure. We examined the discontinuity of the functional sites in the exon-intron structure of genes and the distribution of lengths and phases of the functional site encoding exons in vertebrate genes. It was shown that the DNA fragments coding the functional sites were in the same exons, or neighboring exons. The observed tendency to cluster the exons that code functional sites which could be considered as the unit of protein evolution. We studied the characteristics of the structure of the exon boundaries that code, and do not code, functional sites in Metazoa species mostly presented by Homo sapiens, Mus musculus and Rattus norvegicus. This is accompanied by a reduced frequency of intercodon gaps (phase 0) in exons encoding the amino acid residue functional site, which may be evidence of the existence of evolutionary limitations to the exon shuffling. We projected the known functional sites amino acids positions to the alternative isoforms and examined if the functional sites could change during alternative splicing events. We found only 5 genes of 6000 having sites interrupted in the case of exclusive exon. Also we found that SNP rarely occurred in functional sites and their surroundings. That indicates the high level of conservation of structural-functional organization of proteins. Our results characterize the features of the coding exon-intron structure that affect the functionality of the encoded protein and allow a better understanding of the emergence of biological diversity.

Chromatin changes induced by DNA demethylation

Yulia A. Medvedeva*[1,2], Joaquim Custodio[3], Guerau Fernandez[3], Miquel A. Peinado[3], and Tanya Vavouri[3]

* ju.medvedeva@gmail.com
1 Institute of Biotechnology, Research Center of Bioengineering, RAS, Moscow, Russia
2 Vavilov Institute of General Genetics RAS, Moscow, Russia
3 Institute of Personalized and Predictive Medicine of Cancer, Barcelona, Spain

Methylation is a heritable, yet dynamic, modification of the DNA associated with gene silencing and chromatin compactization. DNA methylation levels are altered dra-

matically in complex human diseases such as diabetes, Alzheimer's and Parkinson's diseases and cancer. Disruption of DNA methylation in various cancers has encouraged a strong interest for DNA demethylating drugs, such as 5-aza-2-deoxycytidine (Decitabine), which is already used in clinic for the treatment of Acute Myeloid Leukemia and Myelodysplastic Syndromes. Its therapeutic activity is thought to be due to hypomethylation and subsequent activation of tumour suppressors. Still, we know little of the consequences of the use of these 'epigenetic drugs' on chromatin organization and gene regulation.

The aim of this work is to understand the chromatin and gene expression changes induced by Decitabine treatment in different genomic contexts. Our work for the first time shows that Decitabine treatment dramatically changes histone modification profiles. We observe loses of active histone marks in promoter regions of actively transcribed genes, while downstream gene regions quite often gain these marks, suggesting possible rearrangments of transcription initiation events. Comparison with DKO (DNMT1/DNMT3a) cell shows that DNA demethylation of the genome affect chromatin even in regions with no methylation changes, suggesting global redistribution of chromatin domains. Our results also suggest that Decitabine alters chromatin structure through both pathways dependent and independent of its DNA methyltransferase blocking activity. Considering the extent of the changes that Decitabine induces throughout the human genome, we propose that its anti-tumour effects are at least partially due to novel mechanisms that act through chromatin instead of DNA methylation.

This work is partially supported by RSF grant 15-14-30002 to YAM.

Short linear motifs derived from fetoplacental proteins: Bioinformatics and molecular dynamics simulation study

N. T. Moldogazieva*[1], I. M. Mokhosoev[1], and A. A. Terentiev[1]

* nmoldogazieva@mail.ru

1 N. I. Pirogov Russian National Research Medical University 111997, 1 Ostrovityanova street, Moscow, Russian Federation

Short linear motifs (SLiMs) have been recognized to be functional important sites in a large variety of regulatory proteins through participating in protein-protein interactions, signal transduction, cell cycle regulation, protein secretion, etc. Both direct (RGD, KSEL, HDEL, PxxP, LxxLL) and reverse (LLxxL) SLiMs have been shown to possess a role in the protein functioning. To date, the functional importance of a definite residue in a protein biologically active sites has been mostly assessed by taking into account physico-chemical properties of amino acid residues. There are only statistical data on the ability of natural amino acids to form a definite secondary structure element and its contribution into conformational changes in peptides and proteins. Earlier, with the use of local sequence alignment, we revealed a stretch of seven amino acid residues, LDSYQCT (residues 14-20), in human alpha-fetoprotein

(AFP), with sequence similarity to a part of receptor-binding site of human epidermal growth factor (EGF) and other growth factors of EGF family, and EGF-like repeats of cell adhesion proteins as well. The sequences were found both in direct and reverse form and were designated as AFP14-20-like motifs. In this work, with the use of bioinformatics tools we extracted from protein primary structure data bases a large variety of proteins that contained direct and reverse AFP14-20-like heptapeptide motifs linked by a consensus octapeptide CxxGY/FxGx sequence. Application of molecular dynamics simulation study allowed assessment conformational/dynamic properties of amino acid residues in short penta- tetra- and heptapeptide linear motifs to reveal structural basis underlying functioning of the peptides. Then, we tested biological activities of some of the peptides to found correlation between conformational/dynamic properties and biological activity of the peptides studied. Additionally, we revealed a structural module composed of 22 amino acid residues, i.e. 22-member module, in a variety of regulatory proteins. Functional annotations allowed classifying these proteins into the following six groups: (i) growth factors, their receptors and intracellular effectors (10%); (ii) cell adhesion proteins and their receptors (14%); (iii) transcription factors (22%); (iv) pro- and antiapoptotic proteins (9%); and (v) enzymes that, mostly, belong to oxidireductases, hydrolases and synthetases (40%), (vi) proteins involved in intracellular protein transportation and localization (5%). Notably, the overwhelming majority of the proteins containing 22-member module may regulate cell proliferation, differentiation, migration and apoptosis. We propose them as structural markers for regulatory proteins that may be involved in embryo- and carcinogenesis.

Novel comparative genomic approach for detecting nonhomologous RNA regulatory elements

Anna Obraztsova*[1] and Zoe Chervontseva[1]

* alcedooo@gmail.com
1 Lomonosov Moscow State University, Russia

Precise coordination of ribosomal protein synthesis is crucial for appropriate assembly of ribosomes. More than half of the ribosomal proteins in *E. coli* are known to be controlled by distinct RNA regulatory elements that occur within their own mRNA. In some cases such regulatory elements mimic the ribosomal RNA site where one or two of regulated proteins may bind providing direct negative feedback from levels of ribosomal proteins. But it seems like this mimicry can be inherent to completely different nonhomologous secondary structures in distinct species. That is why minimal free energy (MFE) based search looks more preferable than methods using evolutionary conservation of RNA secondary structures. In this work we use novel MFE based comparative genomic approach for detecting such nonhomologous RNA regulatory elements and some new predicted ribosomal protein RNA regulators.

Evolution history of lipoxygenase pathway enzymes

Elena Osipova*[1]

* eva-0@mail.ru
1 Kazan Institute of Biochemistry and Biophysics, Kazan Scientific Center, Russian Academy of Science

The evolution of metabolism of taxonomic distanced organisms is associated with changes of signaling pathways including lipoxygenase pathway. The lipoxygenase specificity defines the site of oxygen incorporation fatty acid during hydroperoxides production, designate the courses of following signaling. In plants the hydroperoxy fatty acid is converted by CYP74 family cytochrome P450 enzymes (allene oxide synthase, hydroperoxide lyase or divinyl ether synthase). In other organisms there are CYP74-like enzymes, but they are not observed in all kingdoms. The way of lipoxygenase signaling pathway evolution is not evident. Lipoxygenase and P450 enzymes nucleotides and amino acids sequences were taken from Uniprot.org and Nelson databases. Biological information was taken from PubMed (NCBI), Brenda-enzymes.org, Uniprot.org, Kegg.jp et al. The multiple alignment were performed by MAFFT, ClustalW and Muscle in MEGA. Phylogeny trees were reconstructed using MAFFT and MEGA. We performed the reconstruction of evolution history of lipoxygenase pathway enzymes, which made clear the sequences of lipoxygenases and cytochromes P450 common ancestors. The lipoxygenase specificity variations were provided by duplications of ancestor sequence and subsequent specificity changes. For example, the lipoxygenases of plants are divided into groups according to the specificity catalytic action. The sequence of P450 cytochromes ancestor is analogous to present-day hydroperoxide lyases and allene oxide synthases of CYP74 family of P450. Based on the phylogenetic analysis of biochemically tested lipoxygenases we identified the determinants of substrate specificity (arachidonic or the linolenic acids) of lipoxygenases. It was revealed that the determinants of catalytic specificities of plants lipoxygenases and CYP74 family cytochromes P450 were the unique sites of protein sequences. It was ascertained that catalytic specificity of mammalian lipoxygenases were not depended on the single sites of protein sequences. Moreover, some mammalian lipoxygenases perform the catalytic reactions as plant CYP74 family cytochromes P450 enzymes. Some common properties of lipoxygenase pathway enzymes evolution of organisms from different kingdoms were revealed.

Electrostatics as a new old factor of the natural selection in genome

Alexander Osypov*[1]

* aosypov@gmail.com
1 Institute of cell biophysics of RAS

Genome DNA physical properties define its shape in the functional space and influence its interactions with proteins, esp. for transcription regulation (TR). DNA is

highly charged and electrostatics (E.) contributes greatly to the subject. DEPPDB was developed to provide all available information on these properties of genome DNA combined with its sequence and annotation of biological and structural properties of genome elements and whole genomes, organized on a taxonomical basis. E. potential (EP) is distributed non-uniformly along DNA molecule and correlates, though not exactly, with GC content, strongly depending on the sequence arrangement and its context (flanking regions). Binding frequency of RNA polymerase to DNA along the genome, measured in direct experiment, correlates to the calculated EP. TR areas have EP and other physical properties peculiarities. Binding sites of transcription factors of different protein families in different taxa are located in long areas of high EP. EP distribution on transcription factors protein molecule surface reflects that of their binding sites. Promoters in average have high value and heterogeneity of EP profile. The transcription starting sites of prokaryotic genomes are characterized by extensive (hundreds of bp) zone of high EP and some peculiarities directly around TSS. This is associated with protein binding and formation of physical properties due to transcription machinery. Specific details of the TSS EP architecture are similar in related taxa. Promoters up-element demonstrates electrostatic nature. E. effects on genome functioning interact with other physical properties of DNA, in particular — bending, thermal stability, supercoiling — in both, formation, and TR. The distribution of curved DNA in promoter regions is evolutionarily preserved, and is mainly determined by temperature of habitat. Mesophilic genomes may have different intensity in curvature, while thermophiles and hyperthermophiles lack it overall because of the life under temperature above the curvature-relaxing point that renders this property useless in TR. DNA curvature is related to AT content, strongly curved DNA fragments must possess high A+T content (reverse is not true). There is no decrease in size and prominence of electrostatic deep in extremophyles, proving importance of E. and its differential role vs curvature. EP properties of Mycobacterium leprae TSS reflect massive pseudogenization and strictly intracellular parasitic life with reduced TR. It has smoothed EP profile compared to its close relatives and far less pronounced increase of EP over upstream region where extended EP deep is commonly found. Functional genes possess even less EP typical features, reflecting the diminishing need for extensive TR. E. plays important and universal role in transcription regulation in prokaryotes, affecting proteins binding probability and positioning accuracy. It may influence horizontal gene transfer, TR systems evolution and contribute to genome regulatory regions high AT content in such diverse domains as Bacteria and Archea. Physical properties formation principles affect such fundamental problems as Chargaff's II rule, redundancy of the genetic code, neutrality of synonymous substitutions; and justify the fundamental idea of DNA phenotype, defining the new principle of biophysical bioinformatics.

Methylation and preservation of CpG dinucleotides in human CpG islands

Alexander Panchin*[1]

* alexpanchin@yahoo.com
1 Institute for Information Transmission Problems RAS, Moscow, Russia

CpG dinucleotides are extensively underrepresented in mammalian genomes. It is widely accepted that genome-wide CpG depletion is predominantly caused by an elevated CpG > TpG mutation rate due to frequent cytosine methylation in the CpG context. Meanwhile the CpG content in genomic regions called CpG islands (CGIs) is noticeably higher. This observation is usually explained by lower CpG > TpG substitution rates within CGIs due to reduced cytosine methylation levels. By combining genome-wide data on substitutions and methylation levels in several human cell types we have shown that cytosine methylation in human sperm cells was strongly and consistently associated with increased CpG > TpG substitution rates. In contrast, this correlation was not observed for embryonic stem cells or fibroblasts. Surprisingly, the decreased sperm CpG methylation level was insufficient to explain the reduced CpG > TpG substitution rates in CGIs. While cytosine methylation in human sperm cells is strongly associated with increased CpG > TpG substitution rates, substitution rates are significantly reduced within CGIs even after sperm CpG methylation levels and local GC content are controlled for. Our findings are consistent with strong negative selection preserving methylated CpGs within CGIs including intergenic ones.

Bioinformatic approaches to building of gene regulatory networks

Anna Petukhova*[1], Irina Zhegalova[1], Laura Kazieva[1], Vera Mikhailova[1], and Sergey Suchkov[1]

* annet.057@gmail.com
1 I. M. Sechenov FMSMU, Russia

One of the main tasks of the post-genomic molecular biology and genetics is the researching of principles of organization and functioning of gene regulatory networks — molecular-genetic systems that provide processes of formation of phenotypic features of the organism. The appearance of high-performance methods of molecular biology has resulted in the accumulation of a huge amount of experimental data about structural and functional organization of gene networks and their molecular and genetic components. This information is presented in the large number of publications and computer data bases which are describing different aspects of gene networks functioning. Data about gene networks important for medicine and biotechnology is accumulating especially fast and require instant processing and analysis. Analysis of these experimental data is imposible without using of contemporary information technologies and effective mathematical methods of data analysis and modeling of

biological systems and processes. As a tool that is indispensable for these needs bioinformatics has a wide range of applications for building of gene regulatory networks and also can make this prosses much easier. This review presents basic bioinformatic approaches to building of gene regulatory networks.

StructAlign — a program for alignment of structures of DNA-protein complexes

Yaroslav Popov*[1]

* syav.popoff@yandex.ru
1 Lomonosov Moscow State University, Russia

Comparative analysis of structures of complexes of homologous proteins with DNA is important in the analysis of DNA-protein recognition. Alignment is a necessary stage of the analysis. An alignment is an establishment of equivalences between amino acid residues and nucleotides of one complex and ones of the other. We present the program StructAlign for aligning structures of DNA-protein complexes. The program inputs a pair of complexes of DNA double helix with proteins and outputs an alignment of DNA chains corresponding to the best spatial fit of the protein chains. Currently there are no publications presenting analogous programs. A web interface to StructAlign is available at http://mouse.genebee.msu.ru/tools/StructAlign.html.

The promoter and enhancer landscape of inflammatory bowel disease

Albin Sandelin*[1]

* albin@binf.ku.dk
1 University of Copenhagen, Denmark

Coordinated gene regulation is essential for all aspects of cell biology, including development, differentiation and disease. Characterization of enhancers and promoters in disease has been difficult due to the lack of genome-wide methods suitable for the analysis of small tissue samples. Therefore, we know little about the regulation of genes in disease, and its variation between patients. Related to this, 85% of protein-coding genes show heritable variation in expression due to variance in gene regulation. Thus, localization of promoters and enhancers within patient material is important for disease biology and genetics.

Because promoters and enhancers are transcribed, they can be detected by RNA sequencing. Utilizing this, we have profiled promoter and enhancer usage of the descending colon in 110 patients suffering from inflammatory bowel disease (IBD). To our knowledge, this is the largest study of enhancers in a clinical setting ever done.

IBD is a complex group of chronic inflammatory conditions in the gut. Crohn's disease(CD) and Ulcerative Colitis (UC) are the two principal subtypes. Correct treatment depends on accurate sub-type diagnosis, which is challenging and expensive.

To this end, we identified a promoter/enhancer set that with high accuracy can distinguish the shared inflammatory response, and UC-or CD-specific profiles. We identified >40.000 transcribed enhancer regions, where subsets are specifically induced in general inflammation or in UC/CD. Many of these inflammation-specific enhancer occur in clusters corresponding to so-called super-enhancer regions. IBD-associated SNPs were highly enriched in these regulatory regions, enabling subsequent identification of casual regulatory mutations.

Transcription factors regulating gene expression in different cell types of moss *Physcomitrella Patens*

Tatyana Savelieva*[1]

* savelievatanyya@gmail.com
1 MIPT, Moscow, Russia

Physcomitrella patens is a popular model object in plants systems biology. Notable features of *P. patens* are sequenced genome, high level of homologous recombination among all land plants and existence of various sustainability mechanisms to overcome extreme factors of the environment. Regulation of gene expression in moss *P. patens* under conditions of complex stress is the key to fundamental knowledge, which can be applied in agrobiotechnology. Here, we analyzed data of high-throughput transcriptome sequencing on SOLID 4 (Life Technologies, Applied Biosystems). To detect trancription factors (TFs) correlating with differential expression between protoplasts and protonema cells we used method based on evaluation of TF motif occurences in promoter region with respect to functional analysis. As a result, 15 TFs associated with up-regulated genes in protoplasts and 7 TFs associated with down-regulated genes in protoplasts were identified. According to gene functional analysis part of TFs associated with increased expression in protoplasts regulate genes involved in synthesis of jasmonates and lipids and in biomolecular catabolism. TFs associated with increased expression in protonema, regulate genes responsible for synthesis and decay of structural elements of cells, mostly cytoskeleton. We have not only confirmed part of plant TF functions that were known before, but also showed probable new functions of known TFs. Moreover, gene groups which are regulated by identified TFs were allocated, among them several genes without known function.

FastQ-ome: A random forest ensemble of FastQ reads as decision trees

Praharshit Sharma*[1]

* praharshit@iscb.org
1 Warsaw University of Life Sciences — SGGW, Poland

We take into consideration a sample set of FastQ raw-read data, and compute the "actual" Quality scores (upon deducting the ASCII-offset of +33 or +64 as applicable). From this value, we re-compute the probability of wrong-base-calling, pertaining to PhredQ-score corresponding to each base/nucleotide in the FastQ-file. Then, we compute the Entropy of Information content from this associated probability, taking logarithm to base 4, considering each base to be a "Decision" and build such a "Entropy-Based Decision Tree" for each FastQ-file, noting that we arrive at a "Quad-tree"/Quarternary-tree, which may be rooted/unrooted, with utmost 4 branches at every node. Also, we note that the Decision-tree may be validated using the observed fact that: as we move from root to leaf (or vice-versa), the Quality drops from 5'-end to 3'-end of the single-strand whose raw reads are represented in the FastQ-file. Finally, we generate a "FastQ-ome: A Random-Forest ensemble of FastQ-reads as Decision Trees" and observe some interesting principles/properties of worth to note.

Bayesian Markov models consistently outperform PWMs at predicting motifs in nucleotide sequences

Matthias Siebert[1] and Johannes Soeding*[1]

* soeding@mpibpc.mpg.de
1 Max-Planck-Institute for Biophysical Chemistry, Goettingen, Germany

Position weight matrices (PWMs) are the standard model for DNA and RNA regulatory motifs. In PWMs nucleotide probabilities are independent of nucleotides at other positions. Models that account for dependencies need many parameters and are prone to overfitting. We have developed a Bayesian approach for motif discovery using Markov models in which conditional probabilities of order $k-1$ act as priors for those of order k. This Bayesian Markov model (BMM) training automatically adapts model complexity to the amount of available data. We also derive an EM algorithm for de-novo discovery of enriched motifs. For transcription factor binding, BMMs achieve significantly ($p<0.063$) higher cross-validated partial AUC than PWMs in 97% of 446 ChIP-seq ENCODE datasets and improve performance by 36% on average. BMMs also learn complex multipartite motifs, improving predictions of transcription start sites, polyadenylation sites, bacterial pause sites, and RNA binding sites by 26%–101%. BMMs never performed worse than PWMs. These robust improvements argue in favour of generally replacing PWMs by BMMs.

DNA study of the early Bronze Age humans in the North Caucasus reveals their dual connection with Near East and Central Europe

A. S. Sokolov*[1], A. V. Nedoluzhko*[2], E. S. Boulygina[2], S. V. Tsygankova[2], F. S. Sharko[1], N. M. Gruzdeva[2], A. V. Shishlov[3], A. V. Kolpakova[3], A. D. Rezepkin[4], K. G. Skryabin[1,2,5], and E. B. Prokhortchouk[1,5]

* These authors contributed equally.
sokolovbiotech@gmail.com

1 Institute of Bioengineering, Research Center of Biotechnology of the Russian Academy of Sciences. 33, bld. 2 Leninsky Ave., Moscow 119071, Russia
2 National Research Center "Kurchatov Institute", Kurchatov sq. 1, 123182 Moscow, Russia
3 Novorossiysk Historical Museum, Sovetov Street 58, 353900 Novorossiysk, Russia
4 Institute for the History of Material Culture, Russian Academy of Sciences, Dvortsovaya Naberezhnaya, 18, 191186 Saint-Petersburg, Russia
5 Lomonosov Moscow State University, Faculty of Biology, Leninskie Gory 1-12, 119991 Moscow, Russia

The North Caucasus region is rich in early Bronze Age sites, with burials yielding many artifacts, including those from the Chekon, Natukhaevskaya, Katusvina-Krivitsa kurgan groups (at Krasnodar Krai, Russia) and Klady kurgan (near Novosvobodnaya Village, Republic of Adygea, Russia). According to the mainstream archaeological hypothesis, these sites belong to the Maikop culture (3,700-3,000 years BC), with Novosvobodnaya communities representing an offshoot of Maikop ancestry. However, due to specific differences in Novosvobodnaya artifacts, the Maikop and Novosvobodnaya assemblages could represent two synchronous archaeological cultures living in almost sympatry but showing independent ancestry, from the Near East and Europe respectively. Here, we use target-enrichment coupled with high-throughput sequencing to characterize the complete mitochondrial sequence of three Maikop and Novosvobodnaya individuals. We identified the T2b, N1b1, and V7 haplogroups, all widely spread in Neolithic Europe. We also identify the Paleolithic Eurasian U8b1a2 and M52 haplogroups, which are today particularly frequent in modern South Asia, particularly in modern India. Our data provide a deeper understanding of the diversity of Early Bronze Age North Caucasus communities and hypotheses of its origin. Analyzing non-human sequencing reads for microbial content, we found that one individual from the Klady kurgan was infected by the pathogen Brucella abortus, which is responsible for zoonotic infections from cattle to humans. This is in agreement with Maikop/Novosvobodnaya livestock mostly consisting of domestic pigs and cattle. This paper represents a first mitochondrial genome analysis of Maikop/Novosvobodnaya culture as well as the earliest brucellosis case in archaeological humans.

Genetic structure of *Streptococcus pneumoniae* population

Irina Tsvetkova*[1], Vladimir Gostev[1], Sergey Belanov[1], and Sergey Sidorenko[1]

* i.tsvetik@gmail.com

1 Research Institute of Children's Infectious, Russia

Streptococcus pneumoniae is one of the main pathogens of upper and lower respiratory tract bacterial infections. Pneumococcal diseases prevention is based on immunization with vaccines, which contain from 10 till 23 of the most prevalent capsular polysaccharides from 93 described variants (serotypes). Pneumococcal cohort immunization leads to changes in the population structure of these bacteria. In particular, vaccinal serotypes disappear from circulation and dominant serotypes arise among those strains, which have been considered "rare" before. In some cases, change of the serotype pattern of pneumococcal populations comes out of complex genetic processes, related to horizontal gene transfer. Analysis of the pneumococcal populations structure and monitoring of evolutionary processes, which are the consequence of the pneumococcal immunization, are indispensable for the formation of effective prevention strategy. However, currently applicable methods for typing of pneumococci (Serotyping, Multiple Loci Sequence Typing) have insufficient resolution. The most promising method for typing pneumococci is whole-genome sequencing (WGS). Currently, serogroup 19, represented by two serotypes (19A and 19F), is one of the most prevailing serotypes in Russian Federation. Structure analysis of designated serotypes cps-loci, which contains the genes responsible for the synthesis of capsular polysaccharides, have considerable theoretical and practical interest. Objective: To estimate the structure of *S. pneumoniae* population, represented by serotypes 19F and 19A; to perform the comparative bioinformatic analysis of cps-loci for detection the fact of serotypes switching. Methods Strains, received in Scientific Research Institute of Children's Infections Serotyping was performed by PCR with CDC recommendation. Multilocus sequence typing (MLST) was performed with standard scheme (www.mlst.net). Whole-genome sequencing (WGS) was performed on IonTorrent technique (Life Technologies) / SPN2 isolate and MiSeq (Illumina) / SPN133/3733 isolates. WGS data selection from the published large scale studies and following analysis WGS data (short reads, HiSeq/Illumina) for selected *S. pneumonie* population were obtained from the open database of longstanding large scale study (*Croucher N.J. et al.*, 2013). Quality of the reads and their processing were performed with FASTQC and Trimmomatic softwares. De novo genome assembly was performed with SPAdes v.3.5.0 and SSPACE v.3.0. Reads mapping onto reference genome was performed with Samtools and Bowtie2. Cps loci was annotated using available annotation databases. Results The structure of *S. pneumoniae* population *S. pneumoniae* population (150 isolates), represented by serotypes 19F and 19A, included: 3 isolates, received in Scientific Research Institute of Children's Infections in 2012 from children with acute otitis media and community-acquired pneumonia (2 strains of 19F serotype and one strain of 19A serotype); 41 pneumococci genome sequences from GenBank database (12 strains of 19A serotype and 29 strains of 19F serotype); WGS data (short reads)

for 106 isolates from the open database (*Croucher N.J. et al.*, 2013), received from the child-carriers in 2001, 2004 and 2007 years (10 strains of 19A serotype and 10 strains of 19F serotype, 18 strains of 19A serotype and 16 strains of 19F serotype, 45 strains of 19A serotype and 7 strains of 19F serotype, respectively). All analyzed population included the pneumococcal isolates, circulated within several years from the time of implementation of 7-valent Prevnar vaccine. Genomes for 109 strains (missing in GenBank database) were de novo assemled. The assembly pipeline gave on average a total genome length of 2 161 240 bp with average contig length of 33 191 bp and average N50 of 65 656 bp. To estimate the whole population structure, reads for 109 strains were mapped onto reference genomes *S. pneumoniae*: 19F (Taiwan 19F (PMEN14)) or 19A (TCH8431/19A). Mutations frequency for each serotype was estimated. The number of transitions and transversions for the entire population was ranged from 3518 to 14077 and from 136 to 4746, respectively. The average TS/TV for the entire population was corresponded to 2,6, it was evidence of a directional mutations, which had biological significance. To estimate the population clonality, mapped genome sequences were analysed with BAPS v6.0 statistical software. For cps-loci comparison, corresponding sequences were extracted from de novo assembled genomes and following estimation of recombinations in variable regions was performed. Conclusion Analysis of the genetic structure of the *S. pneumoniae* populations is important for understanding of epidemiology and disease prevention measures planning.

Two methods to calculate P-value of RNA of a definite shape

Denis Vorobyev*[1]

* denis.g.vorobiev@gmail.com
1 Softberry, Inc., Russia

A model of RNA secondary structure is considered which is defined by a given tree topology and a given size range of every element (stem or loop). Additionally, size ranges of the total pattern length or some of its subpatterns can be set. In this work, two methods of calculating the occurrence frequency of such pattern in a random sequence are described. Along with suitably arranged complementary pairs, energy is also taken into account. Different background models of the random sequence are discussed: from the fixed mononucleotide composition (Bernoulli) model and the fixed dinucleotide composition model to more complex ones.

Selection of somatic mutations within transcription factor binding motifs in human cancers

I. E. Vorontsov*[1,2], I. V. Kulakovskiy[1,2], G. Khimulya[1], E. N. Lukianova[1], I. A. Eliseeva[3], D. Nikolaeva[1], V. J. Makeev[1,2,4]

* vorontsov.i.e@gmail.com
1 Vavilov Institute of General Genetics RAS, Moscow, Russia
2 Engelhardt Institute of Molecular Biology RAS, Moscow, Russia
3 Institute of Protein Research RAS, Pushchino, Russia
4 Moscow Institute of Physics and Technology, Dolgoprudny, Russia

Somatic mutations in cancer cells may affect various genomic elements and disrupt important cell functions. In particular, mutations in DNA sites recognized by transcription factors can alter binding affinities and, consequently, modify expression patterns of target genes. Previous studies found positive selection of mutations that change particular binding motifs, i.e., such motifs accumulate more mutations than expected by chance. However, clear evidence for negative selection in regulatory elements remained elusive. We performed bioinformatics analysis of binding affinity alterations induced by cancer somatic mutations in transcription factor binding sites. To avoid statistical bias we account for mutational signatures of different cancer types. In an agreement with previous studies, mutations in several motifs were found under positive selection. However, multiple motifs of FOX, HOX and NR families tend to avoid mutations that lead to significant affinity changes (either gain or loss). Such stability is even more exhibited in DNase accessible regions. These results suggest that purifying selection protects cancer cells from rewiring of regulatory circuits. Further analysis of transcription factors with conserved binding motifs can reveal cell regulatory pathways crucial for the cell survivability and proliferation.

Search of conserved features in protein-DNA complexes via Nucleic acid — Protein Interaction DataBase (NPIDB)

Olga Zanegina*[1,2]

* zanolya@yandex.ru
1 A. N. Belozersky Institute of Physico-Chemical Biology, Russia
2 Lomonosov Moscow State University, Russia

Structures of complexes between proteins and nucleic acids (NA) are widely used in design of molecular biology experiments. Nevertheless, data from a single structure could not always provide information concerning the most important and conserved features of protein-NA interactions. To perform deep and comparative analysis of protein-NA complexes, the Nucleic acid — Protein Interaction DataBase (NPIDB) can be used. NPIDB currently contains 5440 structures of DNA-protein and RNA-protein complexes supported with search and navigation tools. For any structure from NPIDB it is possible to download PDB-structure; find hydrophobic clusters,

hydrogen bonds and potential water bridges in protein-NA complex; view sequences, view 3D structure in Jmol, link to Pfam and SCOP. Original structural classification of DNA-binding modes of protein domains is implemented in NPIDP. For each annotated protein, family superimposition of structures, sequence alignment with secondary structure, conserved interface water clusters and interaction class are available. Based on this data comparative analysis of protein-DNA complexes can be done. Suggested classification deals with the structural protein domains but not with the whole DNA-recognized proteins. It allows predicting potential protein-DNA contacts for structures crystallized without DNA or for proteins which combine several DNA-recognizing domains. Here, using a family of TATA-box binding proteins (TBPs) and LAGLIDADG_1 proteins as an example, we present workflow to analyze protein domains and whole protein complexes with nucleic acids using services and data from NPIDB. The workflow included: (1) creation of structural superimposition and sequence alignment of TBPs and LAGLIDADG-DNA complexes which were generated by PDBeFold service or were obtained from NPIDB; (2) determination of conserved hydrogen bonds and hydrophobic interactions based on structural superimposition, sequence alignment and data of hydrogen bonds and hydrophobic interactions for each TBP and LAGLIDADG-DNA complex; (3) calculation of conserved water molecules clusters with wLake program (for the whole protein) or obtaining of these clusters from NPIDB (for domains); (4) Creation of interaction scheme for analyzed protein-DNA complexes based on the above. Nine amino acid residues making conserved hydrogen bonds, 13 residues participating in formation of two conserved hydrophobic clusters at DNA-protein interface, and four conserved water-mediated contacts were found for TBP proteins. Partial symmetry of conserved contacts reflects quasi-symmetry of TATA-box protein binding structure. In contrast with TBPs, structures of LAGLIDADG proteins do not contain conserved amino acids, but their conserved amino acid positions (14 in total) interact with DNA. While hydrophobic clusters mediate protein interactions with DNA backbone, direct and water-mediated hydrogen bonds form conserved contacts with DNA backbone. Taking into account conserved contacting positions instead of amino acid residues can be useful for homologous proteins with low sequence similarity.