SMTB-2021 Projects

PR01 Evolution of genome complexity: Are bacteria as trivial as we expect? (languages: Russian and English)

Project leaders: Olga Bochkareva, Natalia Dranenko
Team members: Vera Emelianenko, Aygul Nasibullina,
Aleksandr Chistyakov

Rotations available for: GMT+1 - GMT+8 time zones

We used to think that bacteria are well studied, since they are used in various industries. But do we know our tools so well?

Yes, bacteria are simple and even primitive compared to eukaryotic organisms as their genomes usually consist of a single circular chromosome. But if we look closely at the bacterial world, there are a lot of exceptions: secondary replicons, linear chromosomes, the fusion of chromosomes and plasmids, and much more. Sometimes these events occur literally before our eyes due to the plasticity of bacterial genomes.

In our group, we study the evolution of bacterial genomes to understand the patterns and mechanisms behind such events. This year, we are going to focus on the secondary replicons (plasmids, megaplasmids, chromids) and patterns of gene accumulation on different elements of the genome. We will pay special attention to the genes that were duplicated in some strains, as additional gene copies become the basis for formation of genes with new functions. We assume that the presence of additional replicons can reduce the selection pressure on the genome size and create favorable conditions for gene evolution.

PR02 Evolution escape room: How to make the virus great again (languages: Russian and English)

Project leaders: Peter Vlasov, Dmitry Korkin
Team members: Ivanna Ostapchuk, Iakov Bogantsev, Polina Avduinina, Anzhelika Dodonova, Marlen Toktomamatov, Sofia Beliaeva

Rotations available for: GMT-4 - GMT+8 time zones

Understanding protein evolution is one of the pivotal tasks of modern biology. Proteins are the main building blocks of life that carry out the majority of cellular functions. The protein evolution can be viewed as a process of "walking" in an unimaginably huge protein "sequence space". Each point in this space corresponds to a protein sequence, and the transitions between the points (and the distances "covered" during these transitions) correspond to evolutionary events--mutations--substitutions, insertions or deletions of amino acid residues. Of course, many of these points turn out to be physiologically "inappropriate" - due to the inability to fold into a stable structure or perform the function by the corresponding proteins. And the related questions--the ratio of "good" vs "bad" variants, how difficult it is for the evolutionary process to "find" a new suitable state, and how this space is generally arranged--remain open to science. In our project, we would like to study the "sequence space" for some specific viral proteins, since the viruses are known to evolve quickly, avoiding drugs or host immunity or adapting to new hosts. To do so, we propose to combine approaches from two areas of biology: structural and evolutionary.

Preamble: among the effective and widely used antiviral drugs, there are some with mechanisms of action based on the direct inhibition by the drug molecule a functional site of the target protein. The inhibition is carried out through a physical protein-ligand interaction. If this interaction is strong, it leads to undesirable--for the viral life cycle--structural changes. For example, viral envelope proteins cannot assemble into a normal form, they are physically "interfered" by a drug molecule. Because of numerous random mutations, which is typical for viruses, target proteins can acquire variants that generally retain their functionality, but also locally change their structure and significantly weaken (or completely "nullify") the interaction with specific drug molecules. In fact, in this way the virus "escapes" from the inhibitory action of drugs, and can start again to actively "attack" the cells of the body and continue to multiply and proliferate effectively.

In our project, we will select proteins from viruses that are known to cause dangerous infections, and even pandemics, to humans and the antivirals targeting these proteins, for which the molecular mechanism of action is well known, including knowledge of certain mutations affecting the drug efficiency. We will conduct a large-scale computer modeling of a wide variety of potential variants of viral proteins; it would be unfeasible to study these variants experimentally because of the complexity and high cost of such studies. Next, we will simulate a large number of variants (mutants) of viral proteins to understand which of them can retain the ligand binding properties (and thus be neutral) and which ones will become poorly interacting with drugs (and thus be deleterious for this function). We will use the methods of structural bioinformatics to address these questions. Second, we will study the distribution of these mutations in the virtual “sequence space” to find which transitions between the neutral and deleterious variants of the virus are potentially possible. We will eventually seek to understand how difficult (or, on the contrary, simple) it would be for the virus to explore (and exploit) such "trajectories" in order to "escape" from the action of drugs. This part of the project will be done using methods of evolutionary bioinformatics.

PR03-PR05 Ivankov Lab (languages: Russian and English)

Project leader: Dmitry Ivankov
Team members: Marina Pak

Rotations available for: GMT+1 - GMT+8

Projects:

PR03 1. **Protein stability change prediction upon mutation using the model of protein structure vs using protein sequence**

Computational prediction of protein stability change (ddG) upon mutation is an important challenge in structural biology. To date, ddG prediction tools allow estimating the effect of mutation based on protein sequence only. Usage of protein three-dimensional (3D) structure increases the accuracy of ddG predictions; however, the protein experimental 3D structure is not always available. Nevertheless, we can construct the 3D protein structure using modeling methods, such as AlphaFold2. Then, what is a more accurate approach to ddG prediction? To use a sequence-based predictor or to construct a protein model? In this project, we will check the two approaches using a popular sequence- and structure-based predictor DDGun and state-of-the-art protein modeling tool AlphaFold2.

PR04 2. **Playing with AlphaFold2**

Three-dimensional (3D) protein structure prediction from protein sequence for 50 years puzzled scientists. Last year we evidenced a breakthrough: the DeepMind team created the computer program AlphaFold2 which predicts protein 3D structures with experimental accuracy! Three weeks ago the program and its code were released to the public and now we have a unique chance to play with AlphaFold2!

1) **AlphaFold2 and change of protein stability due to mutation.** The main question is as follows: can we use somehow AlphaFold2 predictions to predict the change of protein stability due to point mutations? In this project, we will predict structures of mutant proteins and will try to find metrics of the predictions correlating with experimentally measured values of DDG.

2) **AlphaFold2 and random aminoacid sequences.** But first, we will test AlphaFold2 on a couple of simple tasks. Namely, we will check its predictions for amino acid sequences that for sure are not proteins. To do this, we will make predictions for the sequences which are randomized versions of the amino acid sequence of real proteins.

3) **AlphaFold2 and metamorphic proteins.** Another test example is the class of proteins having simultaneously two stable 3D structures. These proteins are called metamorphic. Currently, only a dozen of such proteins are known to science. Will be AlphaFold2 successful in predicting simultaneously two structures?

4) **AlphaFold2 without multiple sequence alignment.** The most important input data for AlphaFold2 is the so-called multiple sequence alignment (MSA) of the protein at hand. We will try to estimate independently how critical is the presence of MSA for robust predictions. For this, we will find proteins with defined 3D structures but having little or no homologs at all. Unfortunately, such proteins are rare but they exist: for example, some artificial proteins belong to this class.

PR05 3. **Signal peptide interchange**

Signal peptides are short N-terminal sequences in secreted proteins cleaved after a successful transport. It is well known that the evolutionary pressure acting on signal peptides is relaxed compared to the mature proteins. Nevertheless, from the comparison of homologous proteins, we see that sometimes the sequences of signal peptides are much more similar than that of mature proteins. The aim of the project is to figure out what are the reasons for that observation.

PR06 Green Cosmonauts: Predicting Regulatory Pathways Determining the Response to Space Flight Conditions in Plants (languages: Russian and English)

Project leaders: Alexey Doroshkov, Alexander Bobrovskikh

Rotations available for: GMT+1 - GMT+8

Space flight conditions significantly affect the living organisms. In the conditions of the space station, stress factors act on plants, first of all, negligible gravity and various radiation. Due to the need to cultivate food resources during space flight, it becomes necessary to study in detail the mechanisms of plant response to these conditions.

In recent years, an array of data has accumulated on the massive changes in the gene expression during space flight. Hundreds of experiments that have been performed in various conditions - from simulating flight conditions on the ground, suborbital flights, as well as in orbiting space stations. The large variability of the experimental conditions makes it possible to carry out a massive analysis of such data and to establish a connection between the activation of certain genetic systems and the peculiarities of the flight conditions.

PR07 Doubles or brothers? Genetic basis of the evolution of innate immune system cell types (languages: Russian and English)

Project leader: Alexey Doroshkov
Team members: Maksim Deryuzhenko, Elizaveta Silvanovich

Rotations available for: GMT+1 - GMT+8 time zones

One of the basic fundamental principles of multicellular animals organization is the specialization of cells into functional groups or types. During long evolution, the number of cell types is steadily increasing. There are cell types widespread among Bilateria, from roundworms to vertebrates. These are, for example, nerve and muscle cells. These cells show morphological correspondence between evolutionarily distant species. And according to modern data, the genetic regulators of its differentiation also correspond to each other between species. This allows us to consider them as “true” homologues, that is, to assert that these cell types had common ancestors with the same function. However, there are cell types, the evolutionary origin of which is currently shrouded in mystery. These are, for example, the cells of the innate immune system. The history of the study of these cells began with the discovery of phagocytic hemocytes in echinoderms and other invertebrates. Further similar cells were discovered in vertebrates and in very fancy basal multicellular organisms - ctenophores. In this project, we will reconstruct the network of interacting genes that determine the trajectories of innate immunity cells and trace the evolution of key drivers of cell differentiation and try to answer the question “do these morphologically similar cells have a similar genetic nature?”.

PR08 The magnetic magic of manganese in maize (languages: Russian and English)

Project leaders: Ulyana Zubairova, Alexey Doroshkov

Rotations available for: GMT+1 - GMT+8 time zones

Plant hydraulics depends on environmental factors and the structure of the vascular system and affects the functioning and growth of cells. Moreover, plants grown in different climatic conditions differ in size, morphology, and physiological parameters. The project aims to study the interaction of transport and growth processes in plants under stressful environmental factors. We will analyze the data obtained within the experiment in which maize plants were grown under cold, drought, high and low light, and soil salinity conditions. Plant phenotype was recorded daily during exposure to stress factors, and at the end of the experiment, data on transport processes in the leaf growth zone were obtained by magnetic resonance imaging with contrast. Manganese ions were used as a contrast agent. Thus, a general model will link the development of a plant, primarily the processes of cell and tissue growth determined mainly by water transport, and the functioning of the transport system of the xylem vessels, which itself is also formed during growth and morphogenesis.

PR09 Evolution of immune system evasion (languages: Russian and English)

Project leaders: Yuri Wolf, Max Wolf

Rotations available for: GMT-7 - GMT+4 time zones

Evolution of a sequence, folded on a lattice; selection favors changes in the boundary pattern, but is constrained by folding energy.

PR10 Evolutionary conservation of RNA secondary structure (languages: Russian and English)

Project leaders: Yuri Wolf, Max Wolf

Rotations available for: GMT-7 - GMT+4 time zones

Analyze thick nucleotide sequence alignments; find (imperfect) palindromes in the consensus sequence; find if the observed pattern of sequence variation favors the preservation of the palindromic structure relative to structure-neutral expectation.

PR11 Studying regulatory elements in yeast and Dictyostelium discoideum (languages: Russian and English)

Project leader: Irina Zhegalova

Rotations available for: GMT-4 - GMT+4

Regulatory elements are important components of gene expression control and, therefore, are crucial in development. Enhancers are regulatory regions that are located at some distance from transcription start sites and can interact with promoters by forming chromatin loops. Enhancers are characterized by specific histone modifications, which are used by researchers to identify them in the genome. We will consider different approaches used to search for enhancers utilizing data on chromatin state and histone markers. We will also attempt to characterize how these regions interact with promoters.

Another part of the project, which can be done by another team, or, time permitting, by the same team as the first project, will be based on the study of bivalent chromatin. These are regions that simultaneously harbor activating and repressing signals. Previous research has shown that such regions may play an important role during differentiation in mammals. We will try to find these regions in lower eukaryotes, and, if we are successful, will try to characterize their contribution to development in these species

.

PR11a Genome within a genome. Determination of the genomic sequence of the potential bacterial endosymbiont of the Halisarca dujardini sponge (languages: Russian and English)

Project leader: Alexander Cherkasov

Rotations available for: GMT-4 - GMT+4

Contamination of foreign DNA is a common problem when assembling the genomes of various organisms. This is especially true for specimens obtained from the environment, which may contain a significant amount of bacterial symbionts. Such foreign sequences in genomic assemblies are usually removed, but sometimes this is not easy. Methods for decontamination of genomic assemblies are often focused on already known contaminant organisms whose genomes are available in databases. Methods for chromatin conformation capture (in particular, Hi-C) allow us to look at this problem from the other side since they allow one to reliably distinguish the sequences of nuclear DNA from the rest. At the same time, using Hi-C data, it becomes possible to cluster potentially contaminant contigs in draft genomic assemblies into groups corresponding to the genomes of contaminant organisms or symbiont organisms.

The cold-water sponge Halisarca dujardini lives on the coast of the northern seas of Russia. Previous research suggests the presence of bacterial endosymbionts. Using several sets of short and long reads, a preliminary “draft” version of the sponge genome, as well as several Hi-C datasets, an attempt will be made to isolate, assemble and characterize the genome of such a potential endosymbiont.

PR12 Pipeline for long read assembly and mutation identification (languages: Russian and English)

Project leader: Catalin Rusnac
Team members: Aleskei Shevkoplias

Rotations available for: GMT-4 - GMT+8

Our lab’s main goal is to develop a simple process for experimental evolution and make it accessible to everyone. At SMTB 2019 and 2020 we have been working on a low-cost, high-throughput morbidostat which can be used to adapt cells to different stressors, such as antibiotics or heavy metals. The device gradually increases the stress level, which induces selection of fitter genotypes in the continuously evolving population of cells in the vial. In a few days of evolution, mutations that are advantageous under such harsh conditions accumulate in the population. Using whole genome sequencing, we can identify these mutations and understand how the cells adapted to the specific stressor that we applied.

We have nanopore sequencing reads of a few genomes of E.coli and B.subtilis adapted to Nickel or Cobalt metal ion stress. The long read data will allow us to identify relatively large mutations, such as transposon insertions, which have been shown to play a role in adaptive evolution.

The task of the project is to use this preliminary data to develop a data analysis pipeline for high throughput evolution experiments. The long reads have to be assembled into genomes and compared to the ancestral genotype. The output should be in a format that allows the researcher to select mutations of different types and different quality requirements.

PR13 Neural Networks to Generate Synthetic Gene Expression Data (languages: English, Russian)

Project leader: *Laura Aviñó*
Team members: *Vera Terentyeva*

Rotations available for: GMT-7 - GMT+8

Simulating experiments, generating data that looks real, is one of the challenges of actuality. Good simulations would allow scientists to test hypotheses faster and cheaper, understand better biological phenomena, and easily share knowledge with the scientific community.

Here we are going to simulate genetic expression on simple networks. To do so, we are going to use Neural Networks. This Machine Learning algorithm is shown to be able to behave as Logical gates. And Genetic Networks are also shown to behave as Logical Gates. We are going to take advantage of this similarity to build up our new algorithm. So that, we help scientists to build and simulate their own genetic networks.

PR14 Homology of blood cells in insects (languages: Russian and English)

Project leader: Sergey Isaev

Rotations available for: GMT-7 - GMT+8

The diversity of the blood cells in invertebrates is a complicated and rather poorly studied topic. There’s a range of works which consider cytological description of the blood cells population, there's also some research on their biochemical compounds. However, there’s still a lack of understanding of homologies between hemocytes among different invertebrate animals.

Single cell RNA sequencing (scRNA-Seq) is an important NGS technology, rapidly developing in recent years. Using this methodology we can define transcriptomes of hundreds and thousands of single cells in one go, which allows us to correctly identify different cell types and differentiation trajectories in tissues. As for this year, there are 3 organisms for which the blood cells scRNA-Seq experiment was conducted: Anopheles gambiae, Aedes aegypti and Drosophila melanogaster. In this project, we are going to conduct a comparative analysis of blood cell populations of these three insects, and to define transcriptional factors which play a role in the differentiation process of their blood hemocytes populations. In our work we will try to understand, a) if we can outline homogeneous blood cells using only these cells’ transcriptomes, b) how similar hemocyte cell populations are for different dipteras. This work can be a root of understanding of blood cell population homology for other invertebrates as well.

PR15 Time series for DNA-DNA interactions (languages: Russian and English)

Project leaders: Aleksandra Galitsyna, Nikolai Bykov

Rotations available for: GMT-7 - GMT+8

In our lab, we study the fundamental problem in developmental biology. In embryogenesis, the structure of chromatin changes. But the particular changes are understudied. Can we find it based on a graph of DNA-DNA interactions?

How does the node degree, number of edges, and their weight change during embryogenesis? What are other characteristics that one can use?
Detection of clusters in the graphs (HiChew with Nikolai Bykov) and find what might be the reason for the changes.
What changes first, the regulatory state or the structure?
Association of several types of TAD borders with epigenetics for different organisms.

You'll learn the instruments: Python, Jupyter notebook, working with genome and genomic intervals, reading the files with annotations, plotting average TAD/border, and enrichment of annotation, clustering techniques. These techniques are popular in both bioinformatics and biotech, and you will certainly find them a lot if opting for an academic career.

PR16 Gains and loss of genes during adaptive radiation of amphipods in Lake Baikal (languages: Russian and English)

Project leader: Lev Yampolsky
Team members: Larisa Okorokova, Stefan Riegler, Eugenia Pravdolyubova

Rotations available for: GMT-4 - GMT+4

Where do new genes come from? Why and how soon do genes not needed any more disappear? Ideally, to answer these questions, one should look at clades with a rich phylogeny, this was i.e. at species-rich clades, preferably adapted to novel environments. Thus endemic adaptive radiation such as that of Lake Baikal amphipods, are most suitable for these kinds of studies. We will map Vaikal amphibodes' transcriptomes on ancestral reference genomes to detect losses and gains of genes in gene families with functionality relevant to survival in the unique modern and paleo climatic conditions.

PR17 Deep learning in drug discovery (languages: Russian and English)

Project leader: Olga Kalinina
Team members: Ilya Senatorov, Ilya Mazein

Rotations available for: GMT-4 - GMT+8 time zones

Artificial intelligence and machine learning plays an increasingly important role in various biological applications, including drug discovery. In our lab, we will use cutting-edge deep learning methods to dip our toes into this intriguing field. We will train a new model using deep graph neural networks to predict novel small molecules likely to specifically interact with an important class of drug targets.

PR18 Alternative splicing (universe) discovery, or Evaluation of alternative splicing event types distribution in different organisms (languages: Russian and English)

Project leader: Olga Tsoy
Team members: Lyubov Lonishin, Grigory Ryabykh

Rotations available for: GMT-4 - GMT+8 time zones

Welcome to the alternative splicing universe! Eukaryotic genes consist of regions - exons, and introns. Due to alternative splicing exons and introns can create different combinations, and one gene gives rise to many mRNAs and proteins. There are several types of alternative splicing events: exon skipping, intron retention, mutually exclusive exons, etc. But in which organisms are there more skipped exons; and in which are there more retained introns? Might it be that there is no difference between a human, a fly, or Arabidopsis? If there is - could bioinformatic tools give a different answer? And what about the quality of biological data? Let’s try to answer these questions! And master biological data search and analysis, bioinformatic tools installation, even in one’s sleep, and presentation of our results, nicely and honestly.

PR19 Oral carcinoma biomarkers in Taiwanese population (languages: Russian and English)

Project leader: Katerina Nuzhdina
Team members: Alexey Efremov, Yaroslav Lozinsky

Rotations available for: GMT-7 - GMT+8

Comprehensive understanding of human health and diseases requires interpretation of molecular complexity and variations at multiple levels. The development of high-speed analytical techniques such as next-generation sequencing (NGS) have significantly transformed the field of oncology. Approaches that integrate different omics data types have the potential to uncover molecular differences associated with cancer and to improve understanding of the variability in treatment response, two major challenges in oncology.

Our main goal for this project will be to compare and integrate the multi-omics (whole-exome sequencing (WES), RNA sequencing and target sequencing) approach for Oral squamous cell carcinoma patients. We will try to repeat the analysis from paper and verify there findings for survival analysis association.

PR20 Laboratory of functional transcriptomics of intriguing creatures (languages: Russian and English)

Project leader: Oleg Gusev
Team members: Alexander Dekan, Ruslan Deviatiiarov

Rotations available for: GMT-7, GMT+1 - GMT+8

Among the wonderfully numerous forms of life on Earth there are many that have evolved unusual adaptations to extreme environmental conditions. In our laboratory, we will use RNA expression data to study several such species. We will attempt to track the evolutionary fate of new genes and their regulation, which jointly confer adaptations to extreme environments.