SMTB-2020 Projects

Effectors' evolution in pathogenic E.coli: How to Manipulate the Host (languages: Russian and English)

Project leader: Olga Bochkareva
Team members: Aigul Minnegalieva, Yulia Yakovleva, Vera Emelianenko

Rotations available for: GMT-4 - GMT+10 time zones

Shigella, named after a japanese doctor Kiyoshi Shigi, is a genus of bacteria that causes dysentery. Only after 100 years as sequencing technologies were developed it was recognized that the *Shigella* is actually *E.coli* sub-populations that have horizontally acquired the plasmid with the genes of type III secretion system (T3SS). This system allows *Shigella* to deliver its proteins into the host cells to modulate immune response and drives an adaptation to intracellular lifestyle. Such proteins are called effectors and they are encoded on the same pathogenicity plasmid as T3SS. Recently, several such effector genes (E3 ubiquitin ligases) were discovered in chromosomes of *Shigella* species and we have shown that the presence of these particular genes distinguishes *Shigella* from other invasive *E. coli* but their function remains uncharacterized.

Our project will focus on the comparative analysis of the ubiquitin ligases from available genome sequences of *Shigella* strains. Using chromosome maps we will reconstruct the “genomic island” harboring these genes and search for its homologs in genomic and metagenomic data. We will also identify conserved motifs in intergenic regions and will try to predict the mechanisms that regulate the transcription of the chromosome- and plasmid-encoded effector genes.

One more pathovar, enteropathogenic *E. coli* (EPEC), has also evolved via virulence genes acquisition that allow bacteria to bind host’s intestinal cells and elicit an inflammatory response. We will work with newly sequenced atypical EPEC strains and the experimental data about their phenotypic properties. Using comparative genomic approaches we will reconstruct the evolution of genomic pathogenicity islands and find the genes and mutations responsible for different phenotypes.

During our work we will:

- Use methods of comparative genomics to work with sequences, such as homology searches, sequence analysis, structural analysis and phylogenetic analysis;
- Read and analyze scientific literature to search for possible regulators and their binding sites;
- Learn how to use Python to work with data and visualize results;

Our results may lead to a deeper understanding of the mechanisms of pathogenicity and may help in creating novel vaccines.

Laboratory of Rational Drug Design (languages: Russian and English)

Project leader: Peter Vlasov
Team members: Ilya Senatorov, Ilya Mazein,
Kseniya Zaitseva, Sofia Buianova
Polina Avduinina

Rotations available for: GMT-4 - GMT+10 time zones

The main goal of the project is to introduce the participants to modern computational methods used in biomedicine and in the design of novel therapeutics (often referred to as Drug Design). The educational segment includes a discussion of various topics on the interface of molecular biology and applied medicine and the study of potential targets (proteins of interest) for drug design. The research segment of the project will include the study of protein structures and protein-ligand interactions with the use of various biological and medical resources, such as those that house data on genes, genomes, proteins, molecular interactions, various drug data, etc.

We will select several proteins as targets that have known mutations that strongly influence the interaction of these proteins with low molecular compounds (drugs or natural metabolites). The work on the project will be aimed to study (1) how changes on these proteins (caused by mutations in the genes that code them) emerge and are maintained in the course of evolution and how they influence the biochemistry of the organism… and (2) how such changes are influenced by external factors, such as the selection in the course of cancer growth that, for example, reduces the response of these cancers on treatment with various low molecular compounds.

Evolution of the glyoxylate cycle (languages: Russian and English)

Project leader: Fyodor Kondrashov
Team members: Ekaterina Maksimova

Rotations available for: GMT-4 - GMT+10 time zones

We will look at long-term evolution of the glyoxylate cycle focusing on horizontal gene transfer events of two key enzymes of the cycle. We know from a paper that we did about 15 years ago that horizontal gene transfer from bacteria to eukaryotes happened in the past. Presently, much more genomic information is available for us to study the details of when and how the horizontal gene transfer events took place across the tree of life. The work will involve hands-on work with BLAST to identify homologous genes and further work with reconstructing a phylogeny.

Signal peptide exchange (languages: Russian and English)

Project leader: Dmitry Ivankov

Rotations available for: GMT-4 - GMT+10 time zones

Signal peptides are short N-terminal sequences in secreted proteins cleaved after successful transport. It is well known that the evolutionary pressure acting on signal peptides is relaxed compared to the mature proteins. Nevertheless, from 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.

Bacterial Bakery: Inventing a Recipe for Antivirulence (languages: Russian and English)

Project leader: Masha Tutukina, Anna Kaznadzey
Team members: Ivan Randoshkin, Anna Rybina

Rotations available for: GMT-4 - GMT+10 time zones

Escherichia coli is a classic model organism ubiquitously used in micro and molecular biology. It is also a cause of pathogenic infection, causing gastroenteritis, ulcerous colitis, neonatal meningitis and even Crohn's disease. Despite decades of research into this bacteria much remains unknown about the mechanisms of regulation of its pathogenicity and many of their infections are not easily treatable. Over 80% of all chronic infections of the urinary tract are caused by ultrapathegenic E. coli (UPEC). They can form massive biofilms in the urinary tract and because of this they are highly resistant to antibiotic treatment.

To learn how to treat such infections it is necessary to better understand the molecular mechanisms underlying the virulence. Virulence is based on a metabolic change that is regulated on a transcriptional level. Transcription factors, which are special types of proteins, play a key role in these metabolic changes by regulating the expression of various genes.

Many transcription factors can be turned on or off by low molecular compounds - ligands that can bind to the protein, which in turn changes the protein's structure and function. Most ligands are simple sugars or sugar acids. In theory, such sugars can control the virulence of E. coli by preventing them from forming hard-to-treat biofilms. But which sugars are those?

To find an answer to this question we will study transcription factors that are already known to regulate various virulence-related processes. First, we will analyze the RNA sequence data of various strains of E. coli: normal (wild type) strains and strains with deleted genes coding for these transcription factors. We will study what happens to the function of virulence-related genes in various conditions. Therefore, we will identify the most promising transcription factors that provide the most interesting conditions for further experiments. We will then model the 3D structure of these transcription factors and with the help of molecular docking methods predict the ligands that are most likely to beind to them and change their functional properties.

Finally, we will compare our predictions experimentally. We will check if the transcription factors and ligands influence the ability of the E. coli to form biofilms. For the experimental part we will organise real time video from the lab to demonstrate the experiments. We will return to offline computation mode to work on the analysis of the obtained data.

In the course of the project we aim to understand which proteins regulate the virulence of E. coli and determine a detailed regulatory scheme. We will find the most promising sugars that inhibit biofilm formation and may be used as enhancers of antibiotic treatment of E. coli infections.

Ghost Exons (languages: Russian and English)

Project leader: Zoe Chervontseva
Team members: Evgenia Khodzhaeva

Rotations available for: GMT-4 - GMT+10 time zones

Exons are frequently duplicated in eukaryotic genomes. Many of them are annotated as bone fide exons. However, many of them are not present in any annotated genomes and their evolutionary fate is unclear. Such unannotated exons are called “ghost exons”. In the course of the project we will describe their properties and will consider their evolution. We will study if the ghost exons may actually be real exons and transcribed in some organisms by looking for RNA-seq reads that match their sequence. We will also consider which reading frame ghost exons prefer and how frequently they have substitutions.

Community ecology of bacteriophage (languages: Russian and English)

Project leader: Alexander Mikheyev
Team members: Jigyasa Arora

Rotations available for: GMT-7, GMT+1 - GMT+10 time zones

We will be working on host specificity in bacteriophages. First, we'll work through an example of how host specificity is maintained and resistance evolves, using experimental evolution and re-sequencing data. Then we'll embark on a research project, looking at network properties of bacteriophage/host communities from a community ecology perspective.

To See Light without Eyes: Adaptive Evolution of Photoreceptive Systems in Plants (languages: Russian and English)

Project leader: Alexey Doroshkov
Team members: Alexander Bobrovskikh, Elizaveta Silvanovich, Alina Levina

Rotations available for: GMT+1 - GMT+10 time zones

Plants are the primary producers in ecological and engineered systems. Plants are adapted to accept certain characteristics of solar irradiance and departure from natural conditions (for example in artificial light) quickly diminishes biomass production. In addition to nonspecific response to change in lighting conditions (such as change in tissue temperature or molecule ionization), plans also have specific protein receptors that are sensitive to light of various spectra. Understanding the function and evolution of the light receptors to adaptation of extreme conditions - one of important steps in controlling light response in plants.

We will reconstruct the regulatory pathways of specific light receptors in model organisms. We will use functional gene annotation and meta analysis of transcriptome data. We will then identify homologous sequences and study the molecular evolution of the protein components of these pathways: receptors and other factors that encode the signal transfer. We plan to study how the various components of these pathways change over time, such as how the protein structure of receptors changes during the adaptation of plants to extreme conditions. A bonus question of the project consists of using comparative genomics methods to study the signal context in promoters of the genes of interest and determine if these regions are conserved.

Different from Bilateria: regulation of gene expression in Trichoplax adhaerens using context signalling (languages: Russian and English)

Project leader: Alexey Doroshkov
Team members: Maksim Deryuzhenko

Rotations available for: GMT+1 - GMT+10 time zones

Trichoplax adhaerens is a Placozoa, one of the basal metazoan species. These organisms look like amoeba, however, they have at least six cell types with distinct morphology. The exact number of cell types remains unknown. The cell types in T. adherens are very different from the known cell types in other metazoans because T. adherens diverged from other metazoans very early on after the formation of the Metazoa clade. These animals are a unique model to study the evolution of cell types and gene expression regulation. A common method to study cell types is to transform the cells with a plasmid that codes for a fluorescent protein. The plasmid can be constructed in such a way that depending on the sequence of the promoter the researchers can track the identity of specific cell types and track it during development. Chemical transformation of the T. adherens cells that are used to transfect Bilateria is possible, however, the fluorescent protein is not expressed. We hypothesize that this happens because the regulation of gene expression in T. adherens, specifically of the so-called housekeeping genes, is very different from that found in Bilateria.

Our project aims to reval cis-regulatory elements in promoters of housekeeping genes of T. adherens and in groups of celltype-specific genes.

We will analyze transcriptome data from single cell RNA sequencing (scRNA-Seq) and identify clusters of the cell types. On the basis of this analysis we will identify groups of genes that are differentially expressed between the clusters and those that are expressed in all cell types at a relatively high level. We will then perform functional annotation of these groups of genes, create samples for a comparative analysis and analyze differences in the promoter regions of the genes in different groups.

Finally, we will search for known and predicted bridging sites of transcription factors in promoter regions of these genes. Using this information we will design special promoters that will lead to constitutive and specific expression of reporter genes (like GFP) in cells of T. adherens.

Genes vs Patogenes: reconstruction of regulatory pathway response to viral, bacterial and fungal infections of plants (languages: Russian and English)

Project leader: Alexey Doroshkov
Team members: Alexandr Bobrovskikh

Rotations available for: GMT+1 - GMT+10 time zones

Plants experience many different biotic stressors, including diseases that target cells and tissues and physical damage from insect and animal herbivory. However, plants evolved several mechanisms that allow them to respond to these stresses, including in response to pathogens. For example, hypersensitivity of cells to fungal infection allows specific cells to die off and prevent further spread of the infection. Such stress response to biotic and abiotic factors is a complex process that involves multiple connected pathways in the cell. These pathways are regulated by regulated function of groups of genes, which are often called gene networks. Bioinformatics tools are often key to the study of such networks.

Currently, more is known about plant response to abiotic stress than biotic one. However, analysis of existing transcriptome data may reveal new insights to biotic stress response. Specifically, constriction and analysis of gene networks is a promising approach to describe patterns of plant immune response.

We will be doing a meta-analysis of many available transcriptional datasets that were made in the course of experiments on biotic stress response of infected plants. These experiments were carried out to identify genes that participate in plant immune response. We will construct gene networks of plant stress response, integrating data of known immune response genes, and those genes that will be predicted by our meta-analysis. We will then analyze the resulting networks, identify key components and analyze their evolutionary characteristics. In the course of our work the participants will learn the methods of transcriptome analysis, gene network analysis and with modern methods of phylogenetics.

The Dark and the Light Sides of Morphogenesis: a complex analysis of growth in C3 and C4 grasses in conditions of low and high solar irradiation (languages: Russian and English)

Project leader: Ulyana Zubairova

Rotations available for: GMT+1 - GMT+10 time zones

Plant morphogenesis is guided by stimuli of the environment. It is a testament to the wonders of plant plasticity how many different phenotypes can be formed by plants of the same genotype when exposed to various temperatures, humidity or light conditions. The phenotypic differences include changes in size, form, colour, position of leaves and also changes in density and topology of leaf vessels and localization of cells in tissues. During our work we offer to study the changes of a number of parameters as a function of different light intensity on an organismal, tissue and cellular levels. We will use images that were obtained during an experiment on growing wheat and corn in controlled light conditions. We will use the leaf of these grasses to study stress-induced changes in morphogenesis. During development leaves in plants maintain a stationary growth phase for a substantial period of time, which allows us to study sequential changes in morphogenesis by observing changes in cell structure of the leaf. We will use 3D imaging on wheat leaves that were stained with fluorescent markers to study the architecture of the epidermis. This will allow for quantitative morphological characterization of the leaf epidermis. Our results will be used for subsequent development of computer models of morphogenesis that take into account fundamental aspects of the underlying mechanisms of plant growth.

Epidemic modeling (languages: Russian and English)

Project leader: Max Wolf

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

a. Simple Ordinary Differential Equation (ODE) based models: start with SIR, introduce more complex structure (subclasses of categories) and external events (quarantines, migration, etc). Show endemic equilibrium, effects of measures and external disturbances, etc. Learn stopping criteria, visualization, indicators.

b. Demographic ODE models (shifting "age" cohorts): introduce incubation period, symptomatic and asymptomatic transmission etc. Show effects of incubation period and unchecked transmission.

c. Agent-based models: start with SIR, introduce complex behavior. Show stochastic nature, effect of intra-population variation of parameters.

Evolution under conflicting constraints (languages: Russian and English)

Project leader: Max Wolf

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

a. Toroidal space, Moran process, neighbor-biased reproduction. Show short-term symmetry breaking; long-term fixation.

b. Introduce mutual exclusion; make it evolvable. Show trend toward exclusion.

c. Introduce neighbor-spreading infections. Show how trends change.

Skills: advanced programming; intermediate to advanced biology. Resources: programming, preferably with graphics capabilities.

Analysis of coding/non-coding sequences ab initio (languages: Russian and English)

Project leader: Yuri Wolf

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

a. Visualizing ORFs in web-based ORF-finders: convenience and limitations (long/short ORFs. alt codes, etc). Power and limitations of translating searches.

b. Analysis of patterns of variation (1st/2nd/3rd in coding areas) vs homologs at different distances (similarity search, alignment, visualization). Syncod as a statistical test.

c. GC content (and other statistics): period-3 patterns as coding sequence signatures

Modeling short read sequencing and assembly (languages: Russian and English)

Project leader: Yuri Wolf

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

a. Short read sequencing as a random sample from the source sequence; assembly as connected components of 1-d overlap graph. Modeling of requening and assembly; dependence of contig length on read length and read density. Show percolation phase-transition.

b. Introduce non-uniform read initiation; use real sequences to model. Show effects of non-uniformity.

c. Analyze real-life examples of contig length distribution to reverse-engineer read length and read density; estimate genome coverage.

Understanding the role of genetic variation in molecular interactions between coronaviruses and their hosts (languages: Russian and English)

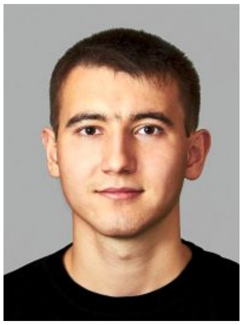
Project leader: Dmitry Korkin
Team members: Rick Kaurov, Anna Culinscaia, Alexander Meister, Anzhelika Dodonova, Oleg Demianchenko, Sofia Belyaeva

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

We are currently leaving in unprecedented times of a coronavirus pandemic. In spite of the whole scientific world joining the forces to fight COVID-19, many questions remain unanswered. In particular, the questions of how mutations in the virus (due to its genetic drift) and in the human host (due to population diversity) affects the molecular interactions between the viral and human proteins. In this project, we will be using structural bioinformatics, sequence analysis, and evolutionary analysis tools to understand the potential impact of the mutations on host-pathogen protein-protein interactions.

Morbidostat software development (languages: Russian and English)

Project leader: Catalin Rusnac



Rotations available for: GMT-4 - GMT+10 time zones

Our lab will work on a morbidostat design that is optimized for scalability. The students' task will be to develop additional software required for running high throughput experiments on the morbidostat. First, additional features will be added to the telegram bot developed last year at SMTB. Second, a PID control algorithm will be implemented for achieving constant OD and growth rate in the morbidostat [similar to https://pubs.acs.org/doi/pdf/10.1021/acssynbio.0c00135]. The students will operate a morbidostat remotely, while developing software that empirically determines the coefficients for a bacterial culture.

Evolution of Tubulins and Microtubules: Microtubules lab (languages: Russian and English)

Project leader: Nikita Gudimchuk
Team members: Alena Korshunova, Anastasia Masaltseva, Iuliia Lopanskaia, Varvara Dreval, Lyubov Makarova

Rotations available for: GMT-4 - GMT+10 time zones

Tubulin is a protein with a central function in the lifecycle of a cell. It forms microtubules that perform a number of critical functions: they determine the location of organelles inside the cell, form the highways for intracellular transport, they form the flagella used in cell motility, they form the machinery that is used in cell division that defines the precision of chromosomal segregation in the dividing cell. An iconic feature of tubulin is its ability to self-assemble into microtubules and then to disassemble. This means that microtubules are elongating and shortening all the time. Tubulin is highly conserved in all eukaryote species, with sequence divergence never being higher than 25% on the amino acid level. Prokaryotes have a protein FtsZ, which participates in cell division, that resembles eukaryotic tubulin in many respects. It is also highly conserved - with sequence identity between eubacteria and archaea being 50%-60%. FtsZ can also form polymers, but it forms linear chains rather than microtubules. FtsZ has a similar tertiary structure to tubulin. However, these two proteins are very different - their sequence identity is only 10%-17%. This level of sequence divergence represents a huge gap between those sequences. How did tubulin evolve at the time of early eukaryote evolution? Which properties are necessary for tubulin to have self assembly into microtubules and subsequent disassembly?

We will use bioinformatics and molecular modeling methods to answer these questions. A hint at how evolution of tubulin could have proceeded may be found in the evolution of other tubulin-like proteins. These are bacterial Btub A/B and other tubulin-like proteins that are only found in some bacterial species. It is possible that the key to microtubule evolution lies in the property of these proteins. With bioinformatics methods we will track the evolution of the sequence and structures of tubulins. Molecular dynamics methods will be used to determine and compare the characteristics of conformational changes of different tubulins. Taken together, these analyses may allow us to understand which key properties must have emerged in tubulins in the course of eukaryote evolution to allow them to perform the wondrous self assembly and disassembly of microtubules.

Generating Synthetic Data (languages: English)

Project leader: Laura Aviñó

Rotations available for: GMT-4 - GMT+10 time zones

Sometimes real data is not enough to perform analysis one may need. For instance, Deep Learning algorithms need a lot of data one may not have, or we may want to simulate specific scenarios that have not happened (at least yet). In this project we will
(1) see different methods to generate synthetic data that "looks" like a real one.
(2) We will download data from the internet (gene expression counts, protein sequences...).
(3) We will learn the key properties of that type of data.
(4) We will generate synthetic datasets that mimic the data we had.
And, finally (5) we will assess how similar or different is our fake data set compared with the real one.

Can we recover the tree of life? (languages: English)

Project leader: Rodrigo Redondo
Team members: Louisa Somermeyer, Stefan Riegler, Arka Pal

Rotations available for: GMT-4 - GMT+10 time zones

We will study the effects of extinction, sampling and Horizontal Gene Transference on phylogenetic inference. We will do computer simulations to generate genes and genomes, compute the phylogenetic tree of these and simulate events of extinction and HGT to test if it is possible to recover the true tree after these events. We will then compute the effects of these events in tree reconstruction using current phylogenetic methods.

Predicting representation of surface proteins using scRNA-Seq expression data (languages: Russian)

Project leader: Sergey Isaev

Rotations available for: GMT-4 - GMT+10 time zones

Cell surface proteins (such as clusters of differentiation proteins, CDs) are important markers of immune cell types; modern classification of leukocytes is based on the representation of these proteins. In the last decade, methods which allow the evaluation of the level of RNA expression in thousands of individual cells (single cell RNA sequencing, scRNA-Seq) have been gaining popularity. As a result, it is necessary to connect the levels of surface proteins representation with the expression of their RNA to correctly determine the cell types under study. The CITE-Seq method which allows to evaluate both of these parameters for individual cells, was created to fill this gap. The main goal of the project is to analyze the relationships between the levels of expression of individual genes and the representation of their protein products on cell surfaces based on several published CITE-Seq experiments.