Санкт-Петербургский государственный университет
Институт биоинформатики
Санкт-Петербургский Академический университет РАН



# Вторая летняя школа по биоинформатике

Санкт-Петербург, 27 июля — 1 августа 2014

*Тезисы докладов*

УДК 004.8
ББК 28.0

*Партнеры:*







Вторая летняя школа по биоинформатике (Санкт-Петербург, 27 июля — 1 августа 2014). *Тезисы докладов.* — Санкт-Петербург: Свое издательство, 2014. — 42 с.

ISBN 978-5-4386-0575-1

# Содержание

# Открытые инструменты для работы с протеомными данными на языке программирования Python

Л.И. Левицкий, М.В. Иванов, А.А. Голобородько, М.В. Горшков

*Институт энергетических проблем химической физики им. Тальрозе РАН 119334, Россия, Москва, Ленинский проспект, д. 38, к. 2; Московский физикотехнический институт (государственный университет) 141700, Россия, Московская область, г. Долгопрудный, Институтский пер., 9 email: lev.levitsky@phystech.edu, тел.: +7 905 758 0818*

В работе представлено описание разработанной нами библиотеки Pyteomics на языке Python с открытым кодом, предназначенной для обработки протеомных данных. Функциональность библиотеки условно делится на три раздела: работа с аминокислотными последовательностями, предсказание свойств пептидов и белков, чтение протеомных данных в распространённых форматах. В рамках этих разделов работа с последовательностями позволяет расчитывать результаты протеолитического гидролиза белков, работать с модификациями, вычислять точные массы и изотопные распределения пептидов и их электрохимические свойства (заряд, изоэлектрическая точка), а также предсказывать хроматографические времёна. Наконец, работа с данными включает в себя чтение файлов в форматах MGF и mzML (результаты LCMS/MS экспериментов), TandemXML, pepXML и mzIdentML (результаты обработки протеомных данных поисковой машиной), и FASTA (базы данных белковых последовательностей). Перечисленные компоненты, а также вспомогательные утилиты для обработки данных (визуализация, регрессия) в совокупности с особенностями языка Python делают Pyteomics доступным и универсальным инструментом для обработки протеомных данных, как в формате интерактивного изучения, так и быстрой разработки сложных приложений.

Примерами применения Pyteomics для написания приложений являются алгоритм валидации пептидных иденти-

фикаций MPscore и протеомная поисковая машина IdentiPy, которые позволяют повысить чувствительность и специфичность протеомного анализа. Более простые примеры утилита для конвертации файлов pyteomics.tandem2xml и приложение psmeval, входящее в состав проекта Galaxy P.

Работа выполнена при поддержке EC FP7 (проект ProtHiSPRA, #282506).

## Resolving variants of unknown significance through reanalysis of 4,978 public RNA-seq samples

D. V. Zhernakova[1#*], P. Deelen[1,2#], M. van der Sijde[1#], J. Karjalainen[1], J. K van der Velde[1,2], M. de Haan[1,2], Kristin M. Abbott[1], C. Wijmenga[1], R. J. Sinke[1^], M. A. Swertz[1,2 ^], J. Fu[1^], L. Franke

*# Contributed equally*
*^ Contributed equally*

*[1] University of Groningen, University Medical Center Groningen, Department of Genetics, Groningen, the Netherlands Oostersingel, 9700RB Groningen*
*[2] Genomics Coordination Center, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands Oostersingel, 9700RB Groningen*
*e-mail: dasha.zhernakova@gmail.com, tel. 89213366329*

In recent years, exome sequencing has emerged as a very effective strategy for genome diagnostics. However, the functional significance is unclear for many of the identified variants, hindering clinical interpretation. To improve upon this, we hypothesized that if a variant of unknown significance is affecting gene expression, it is more likely to be pathogenic (similar to what is known for common disease-associated SNPs, Westra *et al*, Nature Genetics 2013).

We therefore analysed publicly available RNA-seq data from 4,978 human samples from European Nucleotide Archive. We developed methodology to QC and harmonize the RNA-seq data and to account for differences in sequencing strategy, tissue differ-

ences and other (unknown) confounders. We subsequently called SNPs using GATK and imputed genotypes using BEAGLE. We assessed genotype quality using 462 samples for which both RNA-seq data and 1000G genotypes are available (Lappalainen *et al*, Nature 2013) and observed a 97% genotype concordance, indicating that RNA-seq is suitable for genotyping.

This enabled us not only to identify effects of common variants on the gene expression levels of 6,005 genes (*cis*-eQTLs), but also to identify the effects associations of rare variants and gene expression by assessing allele specific expression (ASE). We observed that many rare variants known to be pathogenic strongly associate with gene expression levels.

Since the amount of RNA-seq data that is available in public repositories is growing exponentially, we expect ASE analysis of rare variants will likely provide new tools to resolve many variants of unknown significance.

## GAM: a computational pipeline for integrated transcriptional and metabolic network analysis

Alexey Sergushichev[1,2], Edward Pearce[2], and Maxim N. Artyomov[2]

[1] *Computer Technologies Department, ITMO University, Saint Petersburg, Russia*

[2] *Department of Pathology&Immunology, Washington University in St.Louis, St.Louis, MO*

With an advent of high-throughput transcriptional and metabolic profiling, novel tools for data integration are required. These two types of profiling yield data with complementary characteristics. On one hand, transcriptional profiling is exhaustive, while metabolic profiling is not: all expressed RNAs are identified through RNA-seq, yet absence of metabolite signal does not imply absence of metabolite in the cell. On the other hand, metabolic data are associated with well defined network of mutual biochemical interconversions. Accordingly, transcriptional data inform about expression levels of enzymes regulating metabolite interconversions. Thus, analysis of combined metabolic and tran-

scriptional data in the context of global reaction network would provide integrated view of the data and identify focal points of regulatory network.

In this work we present an R package for network construction and integrated network analysis of transcriptional and metabolic data and a web-service based on this package (https://artyomovlab.wustl.edu/shiny/gam/). Notably, in the case when only one kind of data is available, one can still carry out network analysis based on global reaction network, e.g. providing capability for analysis of transcriptional data in the context of metabolic networks. We have extended approach to analysis of metabolite-enzyme networks adopted by [1] for Milnesium tardigradum.

Implemented network analysis consists of three steps. First, the global network is adjusted to the available data by removing reactions with no expressed enzymes. Second, the network of reactions (including bimolecular reactions) is mapped to a simple graph graph in order to enable existing network analysis algorithms. Finally, the most significantly regulated subnetwork is identified by BioNet [2] algorithm based on p-values.

We applied GAM pipeline to analyze differences between LPS+IFNg-stimulated and unstimulated murine macrophages based on high-throughput metabolic and transcriptional data. We computed most DE module using metabolic and transcriptional data together as well as separately. Notably, difference in regulation of TCA cycle and glycolysis is highlighted on metabolic level, while urea cycle and fatty acid synthesis are best seen on transcriptional level. Yet, only integration of both data types provides a complete picture.

1. Beisser,D., et al. Integrated pathway modules using time-course metabolic profiles and EST data from Milnesium tardigradum. BMC systems biology, 2012, 6(1), 72.

2. Dittrich, M.T. et al. Identifying functional modules in protein–protein interaction networks: an integrated exact approach. Bioinformatics, 2008, 24(13), i223-i231.

# Полногеномный анализ влияния длительного космического полёта на экспрессию мРНК в органах и тканях *Oryzias latipes*

О.С. Козлова

*Казанский (Приволжский) Федеральный Университет*
*42008, Россия, Казань, ул. Кремлёвская, 18*
*e-mail: kazan_capybara@yahoo.com*

Изучение изменений экспрессии в ответ на длительное пребывание в космосе даёт возможность оценить влияние космического полёта на протекание важных биохимических процессов. В данной работе впервые проведена оценка влияния длительного космического полёта на генетическую экспрессию в водном позвоночном, на основе анализа полногеномного профиля экспрессии мРНК (метод mRNA-Seq) в органах и тканях *Oryzias latipes*. В исследовании использовались библиотеки чтений мРНК, выделенной из плавников, сердца, мозга, глаз, мышц, репродуктивных органов и органов пищеварения взрослых рыб, а также мРНК, выделенной из гомогенизированных мальков *Oryzias latipes*.

В каждом образце были найдены группы тканеспецифичных генов, экспрессия которых в космическом полете существенно увеличивалась или уменьшалась (как минимум в 4 раза). Наиболее многочисленные группы генов были обнаружены в образцах из пищеварительного тракта (увеличение и уменьшение экспрессии), сердца (уменьшение экспрессии) и яичника (увеличение экспрессии). Во всех образцах, кроме образца из глаз рыб, было выявлено комплексное изменение экспрессии генов-участников известных биохимических путей, в том числе каскадов, связанных с мышечной и нейрональной активностью организма. Сходная динамика изменения экспрессии одних и тех же функциональных групп генов в разных тканях рыб может свидетельствовать о существовании особой системы регуляции генетической экспрессии (не связанной исключительно с эффектом разгрузки мышечной системы) под действием факторов космического полета.

# Bio4j bioinformatics data platform

P. Pareja-Tobes, A. Alekhin, E. Kovach, M. Manrique, E. Pareja, R. Tobes, E. Pareja-Tobes

*Oh no sequences! Research Group, Era7 bioinformatics*
*Plaza Campo Verde 3 Atico, 18001, Granada, Spain*
*e-mail: aalekhin@ohnosequences.com, +34-693-241-418*

Bio4j (http://bio4j.com) is a high-performance cloud-enabled graph-based bioinformatics data platform. It is specially designed to manage the huge amount of data brought by NGS technologies: it integrates most data available in UniProt KB (SwissProt + Trembl), Gene Ontology (GO), UniRef (50, 90, 100), RefSeq, NCBI taxonomy, and Expasy Enzyme DBs.

Data is organized in a way semantically equivalent to what it represents by taking advantage of the graph structure; in this paradigm it is easy to have many different types of relationships and nodes thus making it perfect for highly interconnected complex biological data. From a performance point of view, relational databases with their tabular data structure are not able to respond to some complex queries that are possible to resolve using the graph paradigm, which gives you fast local access to all the elements related with each entity, through the edges that connect them with others.

Bio4j has a flexible layered structure: on top of it is the abstract domain model interface, then it's implementation using Blueprints, the de-facto standard for graph data modeling, making the domain model independent from the choice of database technology, and in the end there are different technology-specific optimizations for such databases as Neo4j and TitanDB.

The module system based in Statika makes possible to deploy only selected components of the integrated data sets, with Amazon Web Services deployments on hardware specifically configured for them.

Bio4j is open source, available under the AGPLv3 license.

This project is funded in part by the ITN FP7 project INTERCROSSING (Grant 289974).

# Уникальное разнообразие изо-аспартил-метилтрасферазы в насекомом способном к выживанию при полном обезвоживании: геномная организация и экспрессионный ответ на абиотические стрессы

Р.М. Девятияров[1], О.А.Гусев[1], М.Д.Логачева[2] и Т.Кикавада[3]

*[1]Казанский (Приволжский) федеральный университет, 420008, Казань, Россия*
*[2]Московский государственный университет имени М.В.Ломоносова, 119991, Москва, Россия*
*[3]Национальный Институт Агро-биологических Наук, 305-3462, Цукуба, Япония*
*e-mail: ruselusalbus@gmail.com, тел. 89196948678*

Способность личинок африканской хирономиды *Polypedilum vandeplanki* переносить полное высыхание является уникальным примером эволюции защитных систем организмов к экстремальным условиям среды обитания. Анализ генома хирономиды и сравнение его с другими насекомыми показал, что появление способности к ангидробиозу было связано с изменением в геномной структуре и транскрипционной активности ряда групп генов. Одной и таких групп являются белки L-изоаспартат-метилтрансферазы (PIMT), осуществляющие репарацию спонтанно возникающих изо-аминокислот в пептидных цепях. У большинства живых организмов данный фермент представлен единственным геном, тогда как в геноме *P. vanderplanki* был обнаружен участок размером около 30000 пн, содержащий 13 "дополнительных" генов кодирующих PIMT. Классический ген (PIMT1) оказался в другой части генома. Было проведено секвенирование мРНК (методом RNA-seq, Illumina GA IIx) и анализ экспрессионной активности генов с использованием микрочипов в личинках под действием неблагоприятных факторов (тепловой шок, высыхание, радиоактивное облучение). Было обнаружено, что новые гены PIMT показывают значительное повышение экспрессии в ответ на обезвоживание по сравнению с

классическим PIMT1: в 3-40 раз или в 10-230 раз после 24 и 48 часов обезвоживания, соответственно. Кроме того, наблюдаются значительные отличия в динамике изменения экспрессии "новых" PIMT генов в ответ на тепловой шок, радиационное воздействие и окислительный стресс. Учитывая также некоторые отличия в предсказанной структуре белков, можно предположить, что произошло не просто увеличение числа паралогов со сходными свойствами, а имеет место и особая функциональная специализация этого фермента.

## Качественный метод анализа представленности штаммов бифидобактерий в микробиоте кишечника человека на основе генов систем токсин-антитоскин II типа

С.А. Гладышев, К.М. Климина, В.Ю. Макеев, В.Н. Даниленко

*Московский физико-технический институт (государственный университет)*
*Институт общей генетики им. Вавилова РАН*
*e-mail address: gladyshev.sergey@gmail.com*

Метагеномика — одна и интенсивно развивающихся областей человеческой деятельности. Развитие методов секвенирования, применение математических и вычислительных подходов к биологическим задачам послужило основой для развития биоинформатики и системной биологии. Особый интерес приобрел анализ генетических последовательностей, в частности при анализе микробиологических сообществ. Существующие методы метагеномного анализа (на основе 16S рРНК и других генетических маркеров) позволяют анализировать состав микробиологического сообщества только до уровня рода и вида. Но отсутствуют методы, позволяющие проводить детекцию до штаммов. Развитие метода штаммовой идентификации представляет особый интерес ввиду патогенных и пробиотических свойств различтых штаммов.

В работе предложен метод, позволяющий проводить качественную идентификацию штаммов бифидобактерий.

В качестве генетических маркеров были выбраны гены системы токсин-антитоксин II типа (суперсемейства RelBE, MazEF). Поскольку все используемые гены находятся на хромосомах, ожидается достаточный уровень консервативности используемых последовательностей, что делает их пригодными в качестве штамм-специфических маркеров.

В работе предложен алгоритм аннотации генов на основе программ tBlastX и GeneMarkS. Реализованы две вариации этого алгоритма, для аннотации генов в геномах с окончательной сборкой и в геномах с предварительной сборкой. Проверка аннотированных генов проводилась с помощью программы функциональной аннотации аминокислотных последовательностей InterPro.

Разработан метод метагеномного анализа, позволяющий проводить качественную детекцию штаммов по совокупности генетических маркеров. В качестве генетических маркеров выступают варианты генов, аннотированные на предыдущей стадии работы. Для картирования прочтений на генетические маркеры используется программа BowTie2. Постпроцессинг включает в себя подтверждение значимых позиций в выравнивании и определение различимости штаммов на основе покрытия генетических маркеров, построение матрицы различимости штаммов.

Предложенный метод был проверен на ряде образцов микробиоты кишечника, как секвенированных специально для этой цели, так и доступных в других проектах через базы данных NCBI, EBI.

**Литература**

1. Identification and characterization of toxin-antitoxin systems in strains of Lactobacillus rhamnosus isolated from humans. Klimina KM, Kjasova DK, Poluektova EU, Krügel H, Leuschner Y, Saluz HP, Danilenko VN. Anaerobe. 2013 Aug;22:82-9

2. Genetic diversity of the genus Lactobacillus bacteria from the human gastrointestinal microbiome. Botina SG, Koroban NV, Klimina KM, Glazova AA, Zakharevich NV, Zinchenko VV, Danilenko VN. Genetika. 2010 Dec;46(12):1589-97. Russian.

3. Gene identification in prokaryotic genomes, phages, metagenomes, and EST sequences with GeneMarkS suite. Borodovsky M, Lomsadze A. Curr Protoc Bioinformatics. 2011 Sep;Chapter 4:Unit 4.5.1-17.

4. Comparative genomics of defense systems in archaea and bacteria. Makarova KS, Wolf YI, Koonin EV. Nucleic Acids Res. 2013 Apr;41(8):4360-77.

# Поиск молекулярных мишеней для разработки новых противовирусных препоратов в геноме вируса клещевого энцефалита

М. А. Шапиро, М. И. Алейник, А.В. Янцевич

Клещевой энцефалит (КЭ) является одной из самых распространенных и опасных природно-очаговых инфекций лесной зоны умеренных широт. Вирус КЭ способен вызывать у человека острое инфекционное заболевание, характеризующееся лихорадкой, интоксикацией и поражением центральной нервной системы [1,2]. Геном вируса кодирует сравнительно небольшое количество структурных белков и ферментов. Одним из малоизученных ферментов, кодируемых вирусным геномом, является протеазный комплекс NS2B/NS3, состоящий из двух полипептидов и предположительно являющийся сериновой протеазой. Этот ферментативный комплекс участвует в процессинге вирусного полипротеина. В данной работе описывается молекулярное моделирование структуры белка NS2B, входящего в состав протеазного комплекса NS2B/NS3.

## Цель и задачи

Целью данной работы — поиск мишеней для разработки противовирусных препаратов в неструктурной части генома вируса клещевого энцефалита. Для достижения поставленной цели были поставлены следующие задачи:

1. Провести анализ открытых рамок считывания в геноме вируса;

2. Найти фермент, выполняющий важную роль в осуществлении жизненного цикла вируса и который может являться перспективной молекулярной мишенью для противовирусных препаратов; Используя методы *in silico* создать структурные модели данного фермента и провести их анализ;

3. Клонировать гены предполагаемой молекулярной мишени и создать экспрессионный вектор для получения рекомбинантного фермента вируса, который далее будет использован для скрининга ингибиторов *in vitro*.

## Материалы и методы

В проделанной нами работе использовались возможности програмного обеспечения Amber 12-13 [3] на базе вычислительной техники Института биоорганической химии Национальной академии наук Беларуси. Для моделирования использовали два подхода. В первом случае белковая последовательность была разделена на три части по остаткам глицина (как места максимальной гибкости молекулы), для каждого участка аминокислотной цепи проводился фолдинг в течении 100 нс., после чего структурированные фрагменты белка соединялись и проводилась молекулярная динамика длительностью в 370 н.с. Соединение белка проводили путем поворота фрагментов белка, относительно друг друга на один градус и вычисления общего перекрывания атомов двух структур. В итоге выбирался тот угол поворота, при котором взаимное перекрывание было наименьшим.

Во втором случае моделированию подвергалась вся целостная последовательность, но условия моделирования были такими же, как и в первом случае.

Для построения структуры белка по гомологии использовались ресурсы программы Unipro Ugene: gorIV и psipred.

Конструирование экспрессионного вектора pCWori производилось с использованием промежуточной плазмиды pXcmkn12 и использованием рестриктазы XcmI для клонирования в промежуточный вектор и использованием рестриктаз NdeI, XhoI и SalI (рисунок 1).

## Результаты и их обсуждение

Анализ, проведенный с использованием программы Unipro Ugene позволил использовать для оценки структуры два вида алгоритмов предсказания структуры по гомологии — GORIV и PsiPred. Итоги данного прогнозирования противоречат как друг другу, так и результатам анализа с помощью интернет ресурса. Такой результат анализа может объясняться тем, что данный белок принимает свою конечную структуру при взаимодействии с NS3.

После проведенного фолдинга частей белка были получены следующие результаты. Первая часть молекулы (1-44 а.к) при фолдинге приобрела структуру выраженного β — слоя, часть

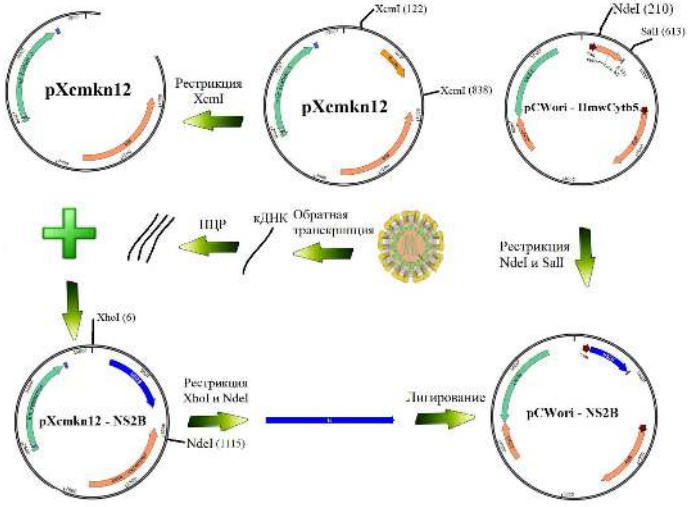



Рис. 1. *Схема получения экспрессионного вектора, содержащего фрагмент, кодирующий последовательность белка NS2B*

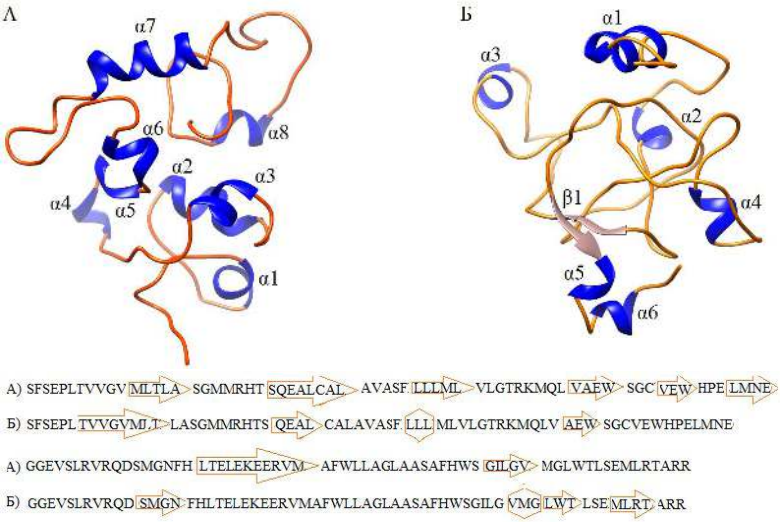
молекулы NS2B с 45 по 82 а.к. не приняла никакой отчетливой конформации, в то же время С-конец белка с 83-ей по 131-ую а.к. богат α–спиральными участками, что может свидетельствовать о его трансмембранной дислокации. После соединения частей белка в одну структуру и последующей динамики был проведён кластерный анализ динамики и выбрана структура с наибольшим временем существования (рисунок 2-А).

На рисунке 2-Б приведен результат динамики целостной структуры, как видно из рисунка, две модели лишь местами напоминают друг друга, а, следовательно, те структурные участки, которые совпали и в первом, и во втором случае — могут присутствовать в нативной структуре белка.

## Заключение

Проведен анализ генома вируса клещевого энцефалита. В качестве предполагаемой молекулярной мишени выбран протеазный комплекс NS2B/NS3.

Результаты структурного анализа белка NS2B указывают на его возможное расположение в фосфолипидной мембране. А также выявлены компоненты, предположительно имеющиеся в нативной структуре белка.




А — результат динамики после соединения разделённого белка. Б — результат моделирования целостной структуры (синим цветом– α-спирали, серым цветом — β-слой). Ниже приведена последовательность NS2B, принадлежащая двум структурам (А и Б) с нанесенными на неё схематическими обозначениями. Стрелка — α-спирали, шестиугольник — β-слой.

Рис. 2. *Результаты проведенных динамик структуры белка NS2B*

Параллельно работам, связанным с исследованием структуры протеазного комплекса, в рамках данной работы проведено клонирование последовательности гена NS2B_437 в экспрессионный вектор pCWori с использованием промежуточного вектора pXcmkn12. Гетерологическая экспрессия в клетках *E. coli* позволит получить препаративные количества данного белка для дальнейших исследований *in vitro*.

**Список использованных источников**

1. Романова Е.В. Сравнительный геномный анализ штаммов вируса клещевого энцефалита, обладающих разной вирулентностью.:Автореф. дисертация на соискание степени кандита биологических наук. Новосибирск, 2011.-17с

2. Belikov S. Activity of recombinant dengue 2 virus NS3 protease in the presence of a truncated NS2B co-factor, small peptide substrates, and inhibitors / Belikov S. // Journal of Biological Chemistry 276, 45762–45771

3. AMBER 12: Reference manual / D.A. Case [и др.]; University of California, San Francisco, 2012. — 348 с.

# Секвенирование и анализ генома *Holospora curviuscula*

Белявская А.Я., Логачева М.Д., Малько Д.Б., Гарушянц С.К., Гельфанд М.С., Раутиан М.С.

Симбиотические бактерии рода *Holospora* — это грам-отрицательные бактерии из порядка *Rickettsiales* класса *Alphaproteobacteria*. Они являются облигатными симбионтами инфузорий из рода *Paramecium*, заселяющими ядра (макро- или микронуклеус) клеток. Эти бактерии имеют сложный жизненный цикл, представленный двумя стадиями: репродуктивной формой, имеющей вид маленьких палочкообразных клеток, и инфекционной — длинными клетками, достигающими 20 мкм в длину. Репродуктивные формы способны бинарно делиться внутри ядра хозяина и давать начало инфекционным формам. Размножаться вне хозяина голоспоры не способны. Все виды *Holospora* обладают хозяинной специфичностью — способностью заражать только определенный вид инфузори, и ядерной специфичностью — способностью заселять только одно из ядер инфузории.

В результате секвенирования (Illumina MiSeq) был получен черновой вариант генома, состоящий из 208 контиг общей длинной 1589519 п.о., средняя длина рида — 212 п.о., содержание G+C пар 37,4%, среднее покрытие — 550.

Автоматическая аннотация генома с помощью сервера RAST выявила 45 генов, кодирующих РНК, и 1548 белок-кодирующих генов, 702 из которых определены как гипотетические белки с неизвестной функцией.

Полученный геном был сравнен с частичными геномами двух других видов *Holospora — H. undulata и H. obtusa*, и 62 полными геномами штаммов 30 эндосимбионтных видов, относящихся к *Alphaproteobacteria* (роды *Anaplasma, Ehrlichia, Neorickettsia, Rickettsia, Orientia, Wolbachia*). Сравнение выявило 433 гена, найденных у всех облигатных эндосимбионтов класса *Alphaproteobacteria*, и 203 гена, характерных только для *Holospora*. Проводится функциональная аннотация генома с целью метаболической реконструкции, а также идентификации генов, отвечающих за специфическое взаимодействие *Holospora* с хозяином.

# Поиск и выделение сайтов инициации репликации (ориджинов) для бактериальных геномов

Выполнил Бутенко Н.А., студент 4 курса,
руководитель д.ф.-м.н. Садовский М.Г.

*ФГАОУ ВПО Сибирский Федеральный Университет ИФБИБТ*
*Miremax0@gmail.com*

Стремительный рост объема генетических данных требует разработки методов их обработки, в частности, выделения семантически значимых сайтов в геномах бактерий на примере ориджинов репликации.

Целью являлась разработка пакета приложений для выделения семантически значимых сайтов в геномах бактерий на примере ориджинов репликации

Были поставлены следующие задачи:

- Создать приложение для обработки генетических последовательностей
- Подобрать генетический материал
- Проанализировать ДНК бактерий

Функция GC-scew — отклонение в разности должных быть парными нуклеотидов G и C. Для определения местоположения ориджина используется минимум этой функции, для точки терминации — максимум.

Первый метод выделения ориджина в области минимума функции GC-scew основан на определении наиболее часто встречающегося олигонуклеотида.

Второй метод позволяет, принимая во внимание *вариативность* ДНК, учитывать возможность замены одного из нуклеотидов в предполагаемом ориджине.

Проведенное сравнение методов получения ориджинов позволяет выдвинуть гипотезу о важности учёта вариативности генома — возможности потери и замены нуклеотида без потери функциональности.

Построена альтернативная классификация и проверка существующей классификации бактерий на основе классификации ориджинов.

**Список литературы**

1. Cambiaire J.-C. de и др. The chloroplast genome sequence of the green alga Leptosira terrestris: multiple losses of the inverted repeat and extensive genome rearrangements within the Trebouxiophyceae. // BMC genomics. 2007. Т. 8. С. 213.
2. Islam M.S. и др. The genome and transcriptome of perennial ryegrass mitochondria. // BMC genomics. 2013. Т. 14. № 1. С. 202.
3. Klasson L., Andersson S.G.E. Strong asymmetric mutation bias in endosymbiont genomes coincide with loss of genes for replication restart pathways. // Molecular biology and evolution. 2006. Т. 23. № 5. С. 1031–9.
4. Nabholz B., Uwimana N., Lartillot N. Reconstructing the phylogenetic history of long-term effective population size and life-history traits using patterns of amino acid replacement in mitochondrial genomes of mammals and birds. // Genome biology and evolution. 2013. Т. 5. № 7. С. 1273–90.
5. Sanderson M.J. и др. Error, bias, and long-branch attraction in data for two chloroplast photosystem genes in seed plants. // Molecular biology and evolution. 2000. Т. 17. № 5. С. 782–97.
6. Sato N. Comparative analysis of the genomes of cyanobacteria and plants. // Genome informatics. International Conference on Genome Informatics. 2002. Т. 13. С. 173–82.
7. Smith D.R. и др. The GC-rich mitochondrial and plastid genomes of the green alga Coccomyxa give insight into the evolution of organelle DNA nucleotide landscape. // PloS one. 2011. Т. 6. № 8. С. e23624.
8. Yura K., Go M. Correlation between amino acid residues converted by RNA editing and functional residues in protein three-dimensional structures in plant organelles. // BMC plant biology. 2008. Т. 8. С. 79.

# Экспрессия генов системы альтернативного сплайсинга, NMD и альтернативных транскриптов гибридного гена RUNX1/RUNX1T1 в клетках ОМЛ

А. В. Войтенкова, О. Д. Кирсанова

## Введение

Острый миелоидный лейкоз (ОМЛ) — злокачественное заболевание крови, характеризующееся экспансией трансформированных белых кровяных клеток. Было показано, что у 20% пациентов малигнизированные клетки содержат транслокацию t(8;21)(q22;q22), результатом которой является образование гибридного гена RUNX1/RUNX1T1, играющего важную роль в развитии и прогрессии заболевания [5].

Данный ген образует необычно высокое количество альтернативных транскриптов (на сегодняшний день идентифицировано 118) [1; 2], подавляющее большинство которых могут образовывать белки, лишенные важных функциональных доменов [7]. Роль таких укороченных транскриптов в настоящее время не ясна [6], однако было показано, что некоторые являются индукторами злокачественной трансформации гемопоэтических клеток [3; 4].

Показанное разнообразие альтернативных транскриптов может быть связано с изменениями в работе системы альтернативного сплайсинга, а также систем, отвечающих за элиминацию нефункциональных транскриптов (например, система уничтожающая мРНК с преждевременными стоп-кодонами — NMD) [8]. Выявление закономерностей образования альтернативных транскриптов гибридного гена может прояснить механизм появления и развития лейкозов.

В связи с этим, целью данного исследования является выявление корреляций в экспрессии альтернативных транскриптов гена RUNX1/RUNX1T1, генов систем альтернативного сплайсинга и NMD.

## Материалы и методы

Из клеток линии Kasumi-1 с помощью набора TRIzol® (Invitrogen, США) была выделена тотальная клеточная РНК. На ее основе была синтезирована кДНК с использованием Oligo-dT и SuperScript III обратной транскриптазы. Уровень экспрессии определялся с помощью количественной ПЦР по отношению к уровню экспрессии гена TBP (TATA Binding Protein).

Результаты корреляционного анализа по Спирмену были использованы для построения теплокарты с помощью функции Reatmap 2 из пакета gplots v.2.13.0 для среды программирования R.

## Результаты и их обсуждение

Для проведения анализа были выбраны ключевые гены систем альтернативного сплайсинга и NMD, экспрессия которых в лейкозных клетках отличается от экспрессии в

нормальных гемопоэтических клетках по данным микроэррей. Из генов системы альтернативного сплай-синга были выбраны следующие: GSPT1,UPF3A, MAGOH, DCP2, DCP1B, SMG1. Гены системы NMD: PTBP1, RBFOX3, RBM25, SRPK2, SRSF6, TIA1.

Альтернативные транскрипты могут возникать как результат использования различных сайтов инициации транскрипции, а также альтернативных сайтов терминации (экзоны 12a, 15a, 17a, 17 гена RUNX1T1). Для определения уровня экспрессии альтернативных транскриптов гена RUNX1/RUNX1T1 использовались пары праймеров к экзонам RUNX1T1: 8a (прямой) + 8b (обратный), 11 (прямой) + 12a (обратный), 15a (прямой) +15a (обратный), 15a (прямой) + 15 (обратный), 16 (пря-мой) + 17 (обратный),16 (прямой) + 17a (обратный). Далее по тексту транскрипты обозначены в соответствии с использованными парами праймеров. Результаты количественного анализа экспрессии альтернативных транскриптов относительно гена TBP представлены в таблице. Следует отметить присутствие в клетках транскриптов 15a_15, несмотря на то, что они являются потенциальной мишенью для генов системы NMD.

*Таблица 1*

**Относительная экспрессия альтернативных транскриптов гибридного гена RUNX1/RUNX1T1**

| Вариант транскрипта | Относительная экспрессия |
|---|---|
| 8a_8b | 0,12 ± 0,03 |
| 11_12a | 0,38 ± 0,06 |
| 15a_15a | 2,34 ± 0,44 |
| 15a_15 | 0,15 ± 0,03 |
| 16_17a | 0,21 ± 0,04 |
| 16_17 | 3,69 ± 0,91 |

Данные, полученные с помощью количественной ПЦР, были использованы для проведения корреляционного анализа по Спирмену. Для визуализации результатов была построена теплокарта (рисунок). Использованный алгоритм позволил не только выявить попарные корреляции, но также выделить кластеры коэкспрессирующихся генов, что отображено на рисунке.

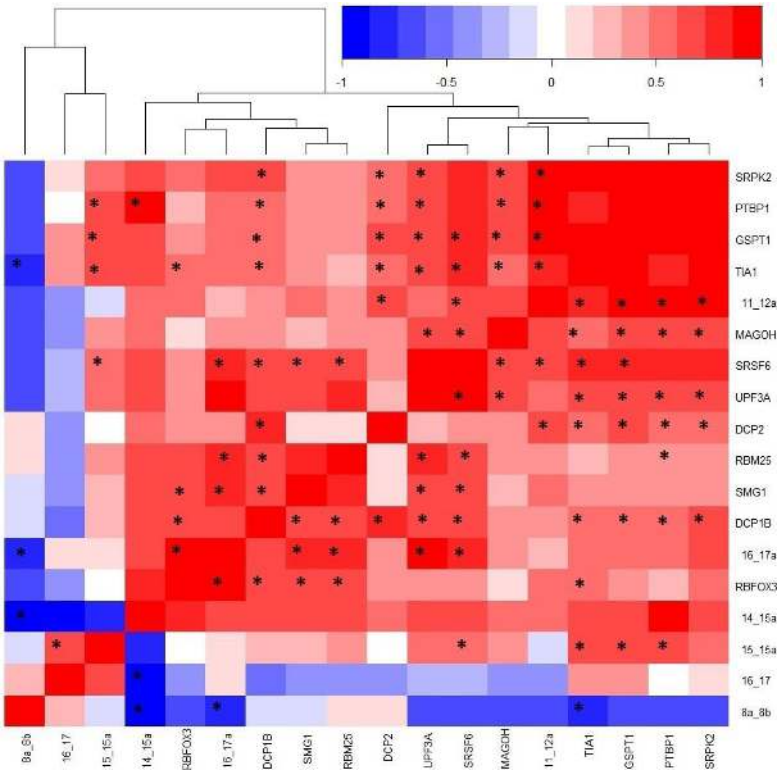



Рис. 1. *Результаты корреляционного анализа по Спирмену*

В верхней части рисунка показаны кластеры коэкспрессирующихся генов. «*» отмечены статистически достоверные корреляции, $p<0,05$. Красный цвет соответствует положительной корреляции, синий цвет — отрицательной корреляции, белый цвет — отсутствию корреляции.

Было выявлено, что экспрессия транскрипта 11_12a достоверно коррелирует с экспрессией 6 генов, среди которых есть как гены системы альтернативного сплайсинга (GSPT1, DCP2), так и гены системы NMD (SRPK2, PTBP1, TIA1, SRSF6). Экспрессия 16_17a транскрипта достоверно коррелирует с экспрессией 5 генов систем альтернативного сплайсинга (UPF3A, SMG1) и NMD (SRSF6, RBM25, RBFOX3). Исходя из различного набора скоррелированных генов, можно предположить различия в регуляции экспрессии соответствующих альтернативных транскриптов.

Экспрессия транскрипта 15a_15 коррелирует с экспрессией 4 генов систем альтернативного сплайсинга (GSPT1) и NMD (PTBP1, TIA1, SRSF6), корреляция с которыми была характерна и для транскрипта 11_12a, что, в свою очередь, говорит о возможной общей регуляции экспрессии соответствующих альтернативных транскриптов. В экспрессии 8a_8b транскрипта были выявлены в основном отрицательные корреляции с другими генами и транскриптами, 3 из которых являются достоверными (включая альтернативные транскрипты 14_15a, 16_17a и ген TIA1 системы NMD).

Таким образом, была доказана и измерена экспрессия альтернативных вариантов транскриптов гибридного гена RUNX1/RUNX1T1 в клетках линии Kasumi-1. Кроме того, были выявлены достоверные корреляции в экспрессии альтернативных транскриптов гена RUNX1/RUNX1T1, генов систем альтернативного сплайсинга и NMD.

В дальнейшем планируется изучить функциональную роль альтернативных транскриптов в развитии ОМЛ, а также роль генов систем альтернативного сплайсинга и NMD в регуляции экспрессии альтернативных транскриптов гибридного онкогена RUNX1/RUNX1T1.

**Литература**

1. Kozu T. et. al. MYND-Less Splice Variants of AML1–MTG8 (RUNX1–CBFA2T1) are Expressed in Leukemia with t(8;21) // Genes, chromosomes and canser. 2005. №43. С. 45–53.

2. LaFiura K. M. et. al. Identification and characterization of novel AML1-ETO fu-sion transcripts in pediatric t(8;21) acute myeloid leukemia: a report from the Children's Oncology Group // Oncogene. 2008. №27. С. 4933–4942.

3. Mannari D. A novel exon in AML1-ETO negatively influences the clonogenic potential of the t(8;21) in acute myeloid leukemia // Leukemia. 2010. №24. С. 891–894.

4. Ming Yan et. al. A previously unidentified alternatively spliced isoform of t(8;21) transcript promotes leukemogenesis // Nature medicine. 2006. Том 12. №8.

5. Miyoshi H. et. al. The t(8;21) translocation in acute myeloid leukemia results in production of an AML 1- MTG8 fusion transcript // The EMBO Journal. 1993. Том 12. №7. С. 2715-2721.

6. Peterson L.F. et al. Acute myeloid leukemia with the 8q22;21q22 translocation: secondary mutational events and alternative t(8;21) transcripts // Blood. 2007. Том 110. С. 799-805.

7. Tighe J.E., Calabi F. Alternative, out-of-frame runt/MTG8 transcripts are encoded by the derivative (8) chromosome in the t(8;21) of acute myeloid leukemia M2 // Blood. 1994. №84. С. 2115-2121.

8. Trcek T. et. al. Temporal and spatial characterization of nonsense-mediated mRNA // Genes Dev. 2013. №27. С. 541-551.

# Могут ли динуклеотидные позиционно-весовые матрицы стать новым «золотым стандартом» для предсказания сайтов связывания факторов транскрипции?

А.В. Денисенко[1,2], И.В. Кулаковский[2,3]

*[1]Федеральное государственное образовательное учреждение высшего профессионального образования «Московский физико-технический институт (государственный университет)» 141700, Московская область, г.Долгопрудный, Институтский переулок, д.9*
*[2]Федеральное государственное бюджетное учреждение науки «Институт общей генетики им. Н.И. Вавилова» Российской академии наук 119991, Москва, ул. Губкина, д. 3*
*[3]Федеральное государственное бюджетное учреждение науки Институт молекулярной биологии им. В.А. Энгельгардта Российской академии наук 119991, Москва, ул. Вавилова, д. 32*
*e-mail: nastia05@me.com*

Распознавание сайтов связывания факторов транскрипции *in silico* имеет ключевое значение в расшифровке регуляции экспрессии генов. Наиболее распространенной вычислительной моделью сайтов связывания является позиционно-весовая матрица, предполагающая независимость позиций в сайтах связывания. Современные экспериментальные методы позволяют получать достаточно данных для построения более сложных моделей. Одна из таких моделей — динуклеотидная весовая матрица- учитывает корреляцию между соседними нуклеотидами. Мы разработали программу для поиска сайтов связывания в последовательности ДНК с использованием существующих динуклеотидных моделей. Высокая вычислительная эффективность поиска и хорошее качество распознавания сайтов связывания, по сравнению с существующими альтернативными подходами (основанными в т.ч. на скрытых цепях Маркова), позволяет выдвинуть динуклеотидные матрицы в качестве нового «золотого стандарта» для компьютерного распознавания сайтов связывания.

## Применение аппарата теории графов к решению задачи о диагностике онкологических заболеваний

А.С. Карсаков

*Нижегородский государственный университет им. Н.И. Лобачевского*
*Россия, 603950, г. Н. Новгород, пр. Гагарина, 23*
*e-mail: karsakov.a.s@gmail.com*

Данный доклад посвящен решению задачи диагностики онкологических заболеваний на основе данных об уровне метилирования ДНК. В работе был применен метод моделирования, основанный на анализе топологии графов. Впервые данный подход был применен в статье [1] для задачи диагностирования заболевания обструктивной нефропатией. В данной работе проведена адаптация процедуры построения графа для поставленной задачи и продемонстрирована адекватность метода на примере решения задачи бинарной классификации для 13 видов онкологических заболеваний. Для большинства видов заболеваний была достигнута точность классификации сравнимая с применением классических методов машинного обучения. Данный подход может быть применен для более широкого класса задач, так как позволяет учитывать связи между изменениями уровня метилирования различных генов.

[1] Zanin M. Boccaletti S. Complex networks analysis of obstructive nephropathy data // Chaos: An Interdisciplinary Journal of Nonlinear Science 21, 033103 (2011).

## Поиск геномных маркеров для определения видов рода *Vibrio*

Керманов А.В.[1], Пшеничный Е.А.[2], Писанов Р.В.[1]

*[1]ФКУЗ Ростовский-на-Дону противочумный институт Роспотребнадзора, г. Ростов-на-Дону,*
*[2]Южный Федеральный университет, г. Ростов-на-Дону*

На настоящий момент насчитывается более 100 видов рода *Vibrio* и этот список постоянно пополняется. Типовым

видом рода является *V.cholerae*, возбудитель холеры. Но еще для десятка видов установлен факт патогенности для человека со способностью вызывать гастроэнтериты или другие клинически выраженные формы кишечных инфекций. Обитая в водной среде и на животных, в пересекающихся экологических нишах, они, очевидно, подвержены масштабному горизонтальному переносу генов, что может свидетельствовать о малоизученности и недооцененности потенциального спектра патогенных для человека вибрионов.

Определение видовой принадлежности вибрионов осуществляется по общепризнанной схеме. Большинство используемых тестов (реакция Фогеса-Проскауэра, тест продукции индола и т.д.) подобраны эмпирически, и для отнесения к конкретному виду необходимы результаты проверки по совокупности 14 биохимических признаков. К главным недостаткам при использовании такого подхода относят нестабильность анализируемых фенотипов и большое количество трудоемких операций по постановке и учету результатов определения.

Предполагается, что задачу видового определения можно решать путем сиквенс-типирования (одно- или мультилокусного). В качестве кандидатов на такие маркеры предлагались гены hsp60, gapA, gyrB, recA, rpoA и pyrH. А С помощью гена *dnaJ*, ранее удачно использованного для дискриминации видов родов Mycobacterium, Legionella и Streptococcus, была успешно определена видовая принадлежность 57 штаммов вибрионов.

In silico-анализ возможных видовых маркеров рода Vibrio был осуществлен Bohle H. M. & Gabaldon T. (2012) , где в качестве маркера был выявлен и экспериментально верифицирован локус VC1988 (ген chromosome segregation ATPase, по штамму N16961). Авторы работы признают, что не все виды надежно дискриминируются даже при использовании данного, подобранного из теоретических соображений, маркера.

Нами предложен подход для оценки дискриминирующей способности маркера исходя из геномных данных, реализована программа на языке Python. Программа позволяет про-

вести расчеты для любого необходимого для исследователя маркера по геномным данным с возможностью выбора локуса, в наибольшей степени отвечающего целям исследования. Предполагается, что выбор маркера будет зависеть от оценки текущей ситуации специалистом, набора патогенов или исследуемых объектов окружающей среды.

Изучается возможность использования гена *hfq* (длина около 260 п.н.) в качестве подобного маркера. Белок, кодируемый данным геном, выступает в роли РНК-шаперона, высоконсервативен внутри вида. Нами был проведен биоинформатический анализ доступных в базе данных Nucleotide генов *hfq* представителей рода Vibrio. Очевидна достаточно надежная дискриминирующая способность маркера даже с использованием выбранной примитивной методики ее оценки. Предстоит подобрать подходящую метрику для рациональной оценки межвидовых вариабельных позиций данного гена и проверить методику экспериментально.

## Metapasta: horizontally scalable tool for microbial community profiling

E. Kovach, A. Alekhin, M. Manrique, P. ParejaTobes, E. Pareja, R. Tobes, E. ParejaTobes

*Oh no sequences! Research Group, Era7 bioinformatics*
*Plaza de Campo Verde, 3, 18001 Granada, Spain*
*email: ekovach@ohnosequences.com, +34 958 25 67 71.*

Metapasta is an opensource, fast and horizontally scalable tool for community profiling based on the analysis of 16S metagenomics data. It is entirely cloudbased and specifically designed to take advantage of it: it performs the community profiling of a sample starting from raw Illumina reads in approximately 1 hour, needing approximately the same time for doing the same on hundreds of samples. It uses BLAST or LAST, but other mapping solutions can be integrated. The taxonomic assignment can be done using a best hit paradigm or a lowest common ancestor paradigm; the user can choose between both assignment algorithms and

setting the similarity parameters required for the assignment. As an output, Metapasta generates the frequencies of all the identified taxa in any of the samples in tab separated value text files. This output includes direct assignment frequencies and cumulative frequencies based on the hierarchical structure of the taxonomy tree. PDF files with assigned taxonomy tree can be rendered. Metapasta is an open source tool available under the AGPLv3 license.

**Methods**

Metapasta is implemented in Scala and based on cloud computing (Amazon Web Services). The graph data platform Bio4j (www.bio4j.com) is used for retrieving taxonomy related information, while Nispero (http://ohnosequences.com/nispero) is used for distributing and coordinating compute tasks.

**Fundings**

This project is funded in part by the ITN FP7 project INTERCROSSING (Grant 289974).

## Поиск мишеней для воздействия на ключевые стадии клеточного цикла Т-клеток на основе конструирования регуляторных сетей Т-клеточного иммунитета человека и мыши

В.И. Конова[1], С.М. Иванов[1], А.А. Лагунин[1,2]

*[1]ФГБУ «ИБМХ» РАМН*
*Москва, ул. Погодинская, д. 10, стр.8*
*[2]ГБОУ ВПО РНИМУ им. Н.И.Пирогова Минздрава России*
*Москва, ул. Островитянова, д. 1*
*e-mail: varvara.konova@ibmc.msk.ru, тел. +7 (499) 255-30-29*

Разработанный ранее метод дихотомического моделирования [1] был применен для создания моделей регуляторных сетей Т-клеточного иммунитета человека и мыши, которые позволили выявить перспективные фармакологические мишени, воздействие на которые может быть использовано для блокирования активации Т-клеточного звена иммунной системы. Регулятор-

ные сети, отражающие процессы передачи сигнала при активации Т-клеточного рецептора лимфоцитов человека и мыши, были построены на основе реакций из коммерчески доступной базы данных TRANSPATH®. Регуляторная сеть для Т-клеток человека, содержала 1618 вершин и 2531 ребро, а для мыши 95 вершин и 121 ребро. В качестве исходных состояний сети были использованы данные о генной экспрессии в Т-клетках, извлеченные из базы данных Gene Expression Atlas (http://www.ebi.ac.uk/gxa). В результате моделирования регуляторных сетей мыши и человека в норме с использованием варианта алгоритма с постоянной активацией гиперэкспрессированных вершин были определены узлы, ингибирование которых можно рассматривать в качестве желаемого события при поиске потенциальных мишеней для пролиферации наивных Т-хелперов мыши. В результате последующего поиска ключевых узлов, блокирование которых приводит к желаемому событию, нами были выявлено несколько десятков одиночных мишеней и парных комбинаций мишеней на основе моделирования человеческой регуляторных сетей. Для мышиной сети была выявлена только одна мишень.

[1] Koborova O.N. et al. SAR and QSAR Environ. Res., 20 (7-8), 755–766.

## *In silico* оптимизация метаболизма *Saccharomyces cerevisiae*

А.В. Котенко, А.С. Розанов, И.Р. Акбердин, С.Е.Пельтек

*Федеральное государственное бюджетное учреждение науки Институт цитологии и генетики Сибирского отделения Российской академии наук (ИЦиГ СО РАН)*
*Адрес: 630090, Новосибирск, Россия, пр.ак.Лаврентьева,10*
*e-mail: kotenkon92@gmail.com*

Переход на использование лигноцеллюлозной биомассы в качестве источника сахаров для промышленной микробиологии требует модификации метаболизма продуцентов с целью активации метаболизма пентасахаров, которые в значительном количестве представлены в биомассе растений. В случае

биоэтанола наиболее эффективным продуцентом являются дрожжи *Saccharomyces cerevisiae*. Для их адаптации к новому источнику сахаров необходимо провести ряд изменений в геноме существующих в настоящее время штаммов-продуцентов. Для определения потенциальных мишеней будет использована математическую модель метаболизма дрожжей *S. cerevisiae*.

К настоящему времени адаптирована математическая модель гликолиза дрожжей *S. cerevisiae*. Она представляет собой систему обыкновенных дифференциальных уравнений, включающих 21 переменную. Каждое дифференциальное уравнения описывает динамику изменения концентрации метаболитов, представленных в исследуемой метаболической системе. В качестве субстрата используется глюкоза, конечный продукт — этанол.

На основе модели был проведен теоретический анализ метаболизма глюкозы в клетке *S. cerevisiae*. Показано, что наиболее перспективными мишенями для оптимизации метаболического пути на увеличение биосинтеза этанола являются гены ферментов алкогольдегидрогеназа, пируватдекарбоксилаза, пируваткиназа, фруктозодифосфатальдолаза и глюкокиназа. Увеличение активности этих ферментов позволяет получить скорость наработки этанола, более чем в 2 раза превышающую таковую в клетках дикого типа.

В дальнейшем планируется развить математическую модель за счет добавления пути утилизации ксилозы.

## Идентификация *hrpL*-зависимых генов *Pectobacterium carotovorum*

С.В. Кузьмич, А.В. Домеников, Е.А. Николайчик

*Белорусский Государственный Университет,*
*биологический факультет, кафедра молекулярной биологии*
*г. Минск, ул. Курчатова, 10*
*тел/факс. +375 (17) 209-58-08 e-mail: kuzmich.sofya@gmail.com*

*Pectobacterium carotovorum* — бактериальный фитопатоген, использующий эффекторные белки для преодоления

иммунитета растения и успешного его заражения. На данный момент охарактеризован только один эффекторный белок DspE, доставляемый в клетки растений при помощи системы секреции третьего типа (ССТТ). Гены ССТТ и эффекторных белков находятся под контролем альтернативного сигма-фактора HrpL, а различный фенотип *hrpL*- и *dspE*-мутантов свидетельствует о наличии дополнительных связанных с патогенностью генов, помимо dspE, регулируемых HrpL. Для выявления таких генов был секвенирован геном P. *carotovorum* 3-2, и в геномной последовательности был проведен поиск HrpL-зависимых промоторов, при помощи программы nhmmer из пакета HMMER 3.1. Такой поиск выявил семь HrpL-зависимых промоторов в пределах *hrp*-кластера генов, присутствие которых ожидалось, а также дополнительный промотор за пределами *hrp*-кластера. Поскольку последний промотор оказался расположен слишком далеко (360 н.п.) от ближашей рамки считывания, с использованием программы TransTerm HP 2.08 было проверено присутствие возможных транскрипционных терминаторов после этого промотора. Один потенциальный терминатор оказался расположен на расстоянии 105 н.п. от промотора, а второй, более стабильный, — на расстоянии 1118 н.п. Остановка транскрипции на первом терминаторе даст транскрипт длиной в 128 н.п., не содержащий открытых рамок считывания. Это дает основания предположить, что данный участок является малой регуляторной РНК (sRNA). С использованием программы IntaRNA был проведен поиск возможных мишеней для этой sRNA в геномах родственных штаммов *P. carotovorum* PCC21 и PC1. Вероятными мишенями для PCC21 являются следующие гены, так или иначе задействованые в патогенезе: *fliJ* (синтез флагеллинового шаперона) и *dppF* (ABC-транспортер олиго пептидов); для PC1 — *fliJ*. Планируется экспериментальная верификация идентифицированного нового HrpL-зависимого промотора и исследование роли находящихся под его контролем генов.

# Клонирование и биоинформатический анализ последовательностей КДНК генов SHP малых гидрофобных белков антарктических мхов

Макаренко Р. А., Моргун Б. В.

На сегодняшний день флора Антарктического региона является малоисследованной. Научный интерес к ней обуславливается наличием адаптаций к негативным климатическим условиям Антарктики (низкие температуры, высокий уровень солнечной радиации, низкая влажность и др.)

Поиск генов, которые отвечают за устойчивость к воздействию окружающей среды, имеет также и практическое значение для улучшения уже существующих свойств сельскохозяйственных растений.

Гены группы SHP кодируют малые гидрофобные белки мембран — класс консервативных протеинов, который встречается в разных группах организмов, включая растения, животных, грибы и бактерии. Было показано, что экспрессия этих белков индуцируется почти всеми видами стрессовых факторов. Основной функцией этих белков является участие в создании и передаче клеточного сигнала в ответ на действие стрессового фактора.

Целью работы было клонирование и молекулярно-генетический анализ последовательностей генов *shp1* и *shp2* малых гидрофобных белков антарктических мхов *Warnstorfia fluitans, Polytrichum juniperinum Bryum pseudotriquetrum* и др. Для этого выделяли тотальную ДНК исследуемых мхов и проводили ПЦР с генспецифическими праймерами, сконструированными на основе данных о последовательностях модельного мха *Physcomitrella patens*. Из семи образцов, у которых проходила реакция амплификации с праймерами, выделяли РНК по методике: гомогенизировали живой материал в буфере для выделения РНК (0,1 М Трис pH 8.0, 5 мМ ЕДТА pH 8.0, 0, 1 М NaCl, 0,5% SDS). Полученный гомогенат очищали фенол-хлороформной смесью, осаждали РНК изопропанолом с 3 М ацетатом натрия и промывали 70 % этанолом. На ма-

тричной РНК с использованием праймеров Oligo (dT)18 проводили обратную транскрипцию — синтез комплементарной ДНК. Целевые гены были амплифицированны при помощи двух пар специфических праймеров SHP1F-SHP1R и SHP2F-SHPR2. Полученные фрагменты кДНК были клонированы в плазмидный вектор pUC19 и перенесены в бактериальные клетки *Escherichia coli* штамм XL- 1 Blue путем химической трансформации. Отбор колоний проводили на среде, которая содержала в качестве селективного агента карбенициллин (100 г/мл). С отобранных колоний выделяли рекомбинантную ДНК и проверяли наличие вставки длинной 200 пар нуклеотидов при помощи рестриктаз HindIII и EcoRI и передавали на секвенирование. Полученные данные анализировали при помощи программы Basic Local Alignment Search Tool (BLAST) на сайте NCBI сравнивая с уже известными нуклеотидными последовательностями Генетического Банка (Genbank).

К настоящему времени были секвенированы фрагменты кДНК генов *shp1* и *shp2* мха *Warnstorfia fluitans*. Анализ последовательностей показал наибольшую степень сродства к следующим видам: *Physcomitrella patens, Ricinus communis, Jatropha curcas, Zea mays, Prunus persica* и др. Сравнение с референтным геномом *P. patens* выявил наличие 3-их синонимических и одной значимой нуклеотидных замен на консервативном участке домена Pmp3 гена *shp1*. По данным литературы отмечено, что домен Pmp3 непосредственно участвует в создании внутриклеточного сигнала через активацию MAP-киназного пути. Для гена *shp2* было обнаружено 11 значимых замен на разных участках кДНК. Данные отличия могут иметь значения в регуляции активности малых гидрофобных белков и оказывать комплексное влияние на устойчивость исследуемых растений. Дальнейшие исследования, в частности анализ и сравнение результатов секвенирования других образцов, позволят более полно ответить на вопрос о роли белков семейства SHP в механизмах защиты растений от действий абиотических стрессов, определить филогенетическое родство между различными представителями флоры Антарктического региона.

# Plastid genomes in non-photosynthetic orchids: sequencing, assembly and analysis of gene content and evolution

M.V. Matveeva[1], M.D. Logacheva[2]

*[1]KAZAN FEDERAL UNIVERSITY, Kazan, Russia, Kremlevskaya str. 18, 420008 E-mail: rinne.swan@gmail.com*
*Mob. Phone: +79600383881*

*[2]M.V. LOMONOSOV MOSCOW STATE UNIVERSITY, Moscow, Russia, Leninskie gory 1/73, 119991*
*E-mail: maria.log@gmail.com*

Ability to photosynthesis is one of principal characteristics of plants. Organelles of symbiotic origin, called plastids, play a key role in this process. Plastids have their own genome that encodes part of photosynthesis, transcription and translation related proteins. The loss of photosynthetic activity that occurs in several plant lineages causes irreversible change of nutrition type from autotrophic to heterotrophic. This apparently leads to the relaxation of selection acting on plastid genes and could potentially lead to complete elimination of plastid genome. In spite of this all non-photosynthetic plants retain their plastid genomes, though usually they are reduced. Ribosomal RNA genes, as well as several transfer RNA and ribosomal protein genes are universally present in all plastid genomes, even highly reduced, suggesting that translation takes place in plastids of non-photosynthetic plants. This evidences that some plastid genes have additional functions besides photosynthesis; these functions are by now unclear and can be revealed by comparative genome analysis.

A good model for studying patterns and constraints of plastid genome evolution in non-photosynthetic plants are orchids, a large and diverse plant family that contains many cases of independent transition to heterotrophy. We sequenced and assembled plastid genome of *Cephalanthera exigua*, a mycoheterotrophic orchid with rather recent transition to heterotrophy. Comparison of its structure and gene content with that of its photosynthetic relatives, on the one hand, and with other non-photosynthetic plants, on the other hand, sheds light on early stages of plastid genome reduction.

# Поиск генов протеаз в геноме *Bacillus pumilus* 3-19

Митрофанова О.С., Тойменцева А.А., Шарипова М.Р.

*Казанский (Приволжский) Федеральный Университет*
*E-mail: olka29.09.1994@mail.ru*

Штамм бактерий *Bacillus pumilus* 3-19 известен как продуцент гидролитических ферментов — щелочной рибонуклеазы, фосфатазы, субтилизиноподобной протеиназы, глутамилэндопептидазы, новой секретируемой адамализиноподобной металлоэндопептидазы (первый бактериальный гомолог эукариотических адамализинов). В отличие от своего природного предшественника (штамма *B. pumilus* 7P) штамм 3-19 вместе с приобретением устойчивости к антибиотику стрептомицину приобрел способность к повышенному синтезу некоторых гидролаз. Так, продукция щелочной рибонуклеазы увеличилась в 100 раз.

С целью анализа секреторного потенциала бактерий *B. pumilus* 3-19 было проведено полногеномное секвенирование штамма на платформе Ion Torrent. В результате геном штамма *B. pumilus* 3-19 был собран в 39 контигов с общей длиной 3,576,473 п.о. Сравнение контигов штамма с базой данных нуклеотидных последовательностей EMBL показало высокую степень сходства анализируемого штамма с геномом *B. pumilus* SAFR-032 (идентичность составила 93%). Аннотация генома была проведена с использованием on-line алгоритма RAST (rast.nmpdr.org). RAST-анализ позволил выявить 30 потенциальных генов относящихся к системам деградации белков. В отличие от аннотированного штамма *B. pumilus* SAFR- 032, исследуемый штамм 3-19 содержит больше АТФ зависимых протеаз. RAST-анализ не выявил не одного гена секреторной протеазы. Анализ генома *B. pumilus* 3-19 в программе MEROPS (merops.sanger.ac.uk) позволил найти 160 генов пептидаз, в числе которых более 12 потенциальных генов внеклеточных гидролаз.

# Новый MDR метод выявления факторов. ассоциированных с компексными заболеваниями

Ракитько А. С.

*Московский государственный университет имени М. В. Ломоносова, Механико-математический факультет, Москва, Россия*
*E-mail: rakitko@gmail.com*

Во многих стохастических моделях возникают данные высоких размерностей. Подобная ситуация характерна для медико-биологических исследований, в которых функция отклика $Y$ , описывающая состояние здоровья пациента, зависит от набора факторов $X=(X_1,\ldots,X_n)$. Так, например, $Y=1$ или $Y=-1$ означает, что пациент болен или здоров соответственно. В генетических задачах в качестве факторов рассматриваются одиночные нуклеотидные полиморфизмы (SNP), а также факторы окружающей среды (кровяное давление, степень ожирения и т.д.). Однако, зачастую функция отклика $Y$ зависит не от всех факторов $X=(X_1,\ldots,X_n)$, а лишь от некоторого *значимого* набора $\{X_{k1},\ldots,X_{kr}\}$, где $1 \leq k_1 < \ldots < k_r < n$. Выявление значимых наборов представляет собой нетривиальную проблему, которой посвящена целая область научных исследований (GWAS).

Для решения подобных задач в [2] был разработан метод, основанный на понижении размерностей. Основная идея заключается в том, что функции $f$, прогнозирующие отклик $Y$ по различным наборам факторов, упорядочиваются в соответствии с некоторым функционалом ошибки $Err(f)$ для произвольных штрафных функций $\psi$. Распределение случайных элементов $X$ и $Y$ неизвестно, поэтому статистические выводы основаны на оценках $\widehat{Err}_K(f_{PA})$функционала ошибки, вовлекающих предсказательный алгоритм $f_{PA}$, оценку штрафной функции $\hat{\psi}$ и $K$-кросс валидацию.

В статье [1] описанная выше задача была обобщена на случай многозначной функции отклика. Также рассматриваются результаты применения построенной процедуры к анализу сгенерированных данных и данных, полученных клиническими исследованиями.

**Литература**

1. А.В. Булинский, А.С. Ракитько, Оценивание небинарного случайного отклика, Доклады РАН, 2014, т. 455, №6, с. 1-5.

2. A. Bulinski, O. Butkovsky, V. Sadovnichy, A. Shashkin, P. Yaskov, A. Balatskiy, L. Samokhodskaya and V. Tkachuk, Statistical Methods of SNP Data Analysis and Applications, Open Journal of Statistics, Vol. 2 No. 1, 2012, pp. 73-87. A. Bulinski, O. Butkovsky, V. Sadovnichy, A. Shashkin, P. Yaskov, A. Balatskiy, L. Samokhodskaya and V. Tkachuk, Statistical Methods of SNP Data Analysis and Applications, Open Journal of Statistics, Vol. 2 No. 1, 2012, pp. 73-87.

# Использование средств информатики для определения положения биологически активных точек

А.Ю. Чернышёва

*Харьковский Национальный Аэрокосмический университет им. Н.Е Жуковского «ХАИ» 61070, Украина, г. Харьков, ул. Чкалова 17*
*e-mail: AnnaChernyshova1@yandex.ua, тел: 095-55-31-758*

Рассматривается способ повышения эффективности обработки диагностических данных, полученных методом электропунктурной диагностики (ЭПД). Достоверно зная диагностические параметры проводимости в ряде биологически активных точек (БАТ), можно поставить диагноз о заболевании конкретного органа, либо организма в целом. Возникает возможность осуществления мониторинга уровня заболеваний.

Предлагается матричная конструкция электрода (3х3), а также эффективный алгоритм программной обработки данных для определения поверхностного распределения электрических сопротивлений в БАТ. Анализ поверхностного распределения сопротивления позволит оперативно определить местоположение БАТ и вычислить виртуальную точку экстремума.

Для обработки информации массива данных применен метод весовой аппроксимации. В качестве аппроксимирующей функции был выбран многочлен второй степени P(x,y). В качестве весовых функций была выбрана потенциальная функция Коши. Были получены выражения для определения коэффициентов регрессии:

$$a_{00} = P_2(\bar{x},\bar{y}) = \frac{1}{15}\left(2\cdot(z_{12}+z_{21}+z_{23}+z_{32}) - z_{11} - z_{13} - z_{31} - z_{33} + 11\cdot z_{22}\right);$$

$$a_{10} = \frac{\partial}{\partial x}P_2(\bar{x},\bar{y}) = \frac{3}{7}\left(\frac{1}{2}(z_{32}-z_{12}) + \frac{1}{3}(z_{31}-z_{11}+z_{33}-z_{13})\right);$$

$$a_{01} = \frac{\partial}{\partial y}P_2(\bar{x},\bar{y}) = \frac{3}{7}\left(\frac{1}{2}(z_{23}-z_{21}) + \frac{1}{3}(z_{33}-z_{31}+z_{13}-z_{11})\right);$$

где zij — результаты измерения в заданных точках.

Полученные новые выражения обобщают одномерные аналоги. Вывод значения коэффициентов на монитор позволяет пользователю уточнять положение точки БАТ путем простого передвижения по поверхности кожи. Дополнительно определяются координаты нахождения точки экстремума. В результате были найдены числовые значения коэффициентов регрессионной зависимости, а также выражения для нахождения координат точки экстремума. В дальнейшем предполагается оценить эффективность применения матричного электрода, а также провести моделирование работы предложенного алгоритма обработки данных.

## Поиск корреляции между социально-экономическим статусом и профилем метилирования в геноме человека

Григорьев К. А.
Научный руководитель: Добрынин П.

*Институт биоинформатики*
*Центр геномной биоинформатики СПбГУ*

Метилирование ДНК является биологическим процессом модификации молекул ДНК без изменения нулкотидной последовательности, приводящим к изменению уровня экспрессии генов в клетках организма.

Корреляция между влиянием окружающей среды, диеты и прочих факторов и эпигенетическим профилем (в том числе, метилированием CpG-мотивов) в настоящее время является предметом обширного изучения.

С целью установить, существует ли зависимость профиля метилирования от факта усыновления, был проведен сравнительный анализ метиломов российских сирот, усыновленных резидентами США, соответственно, до и после усыновления (временной промежуток 6 месяцев).

Профили метилирования исследуемых индивидуумов были получены с помощью метода MBD-Seq.

В ходе исследования разработан пайплайн, позволяющий оценить различия в профилях метилирования до и после усыновления, выявить гены, на экспрессию которых эти различия влияют, и кластеризовать найденные гены по их функции. На тестовых образцах были обнаружены изменения в метилировании ряда генов, в частности, ответственных за развитие организма.

Планируется применять разработанный пайплайн в дальнейших исследованиях и провести биологическую интерпретацию данных, полученных на расширенном датасете.

## Проект по аннотации генома кубинского попугая *Amazona leucocephalia*

С. Колчанова
Научный руководитель: Павел Добрынин

*Центр геномной биоинформатики им. Ф.Г.Добржанского СПбГУ*
*Институт Биоинформатики СПб*

Кубинский амазон (*Amazona leucocephala*) — это вид с пятью подвидами, разрозненные популяции которого разбросаны по территории Кубы, Багамских и Каймановых островов. Среды обитания их изолированы друг от друга и значительно варьируют по условиям существования для животных. В результате подвиды имеют различные ареалы и диету, уникальные паттерны окраски оперения. Эти попугаи находятся под угрозой исчезновения из-за уничтожения естественной среды обитания и интенсивного отлова для содержания в неволе.

Геном этой птицы был собран, но пока не проаннотирован. Информация о детальной структуре и функционировании генома может быть использована для картирования признаков (например, особых паттренов окраски оперения), анализа эволюции генов и генных семейств у птиц, в филогенетических и популяционных исследованиях, а также помочь в сохранении исчезающих видов.

В рамках данного проекта были проаннотированы гены с использованием референса и поиска гомологии и de novo, а также была произведена коррекция найденных открытых рамок считывания. Проаннотированы повторы с использованием различных программ маскирования, подсчитаны проценты различных типов повторов. Наконец, были проаннотированы SNV и осуществлена попытка предсказать, какие эффекты найденные варианты могут оказывать на известные гены.

Мы планируем, помимо уже сделанного, расширить и уточнить аннотацию генов на новой версии сборки с использованием других референсных геномов и методов.

**Вторая летняя школа по биоинформатике**
Санкт-Петербург, 27 июля — 1 августа 2014
*Тезисы докладов*

ООО «Свое издательство»
19004, Санкт-Петербург, 1-я линия В. О., 42
Тел.: (812) 612-18-81
isvoe.ru editor@isvoe.ru
Подписано в печать: 25.07.2014.
Печать цифровая. Тираж 100 экз.