УДК 577+004.42

**Информационно-вычислительная инфраструктура для поддержки метагеномного анализа**

Морозов А.А.1, 2\*, Шигаров А.О.2, 3, Малков Ф.С. 2, 3, Паскал К.К. 2, 3,

Черкашин Е.А. 2, 3, Лихошвай Е.В.1

1 Лимнологический институт СО РАН, г. Иркутск

2Иркутский научный центр СО РАН, г. Иркутск

3Институт динамики систем и теории управления им. В.М. Матросова СО РАН, г. Иркутск

morozov@lin.irk.ru

Метагеномный анализ, хотя и позволяет описывать микробные сообщества с ранее недоступной точностью, требует значительных вычислительных мощностей и участия специалиста-биоинформатика. Это осложняет его внедрение в широкую биологическую практику. Нами начата разработка удобной для неспециалиста системы хранения и анализа метагеномных данных, которая позволит предоставить исследователям удобный интерфейс анализа и обеспечить хранение в унифицированной форме исходных данных и результатов анализа. В данной статье представлены используемые в проекте система программирования потоков данных и схема SQL-базы данных, используемой для хранения метаданных.

Ключевые слова: биоинформатика, метагеномный анализ, Big Data

**ВВЕДЕНИЕ**

В последнее десятилетие в результате изобретения методов секвенирования нового поколения (NGS, Next Generation Sequencing) и внедрение их в практику исследований биологических систем возник такой раздел молекулярной генетики, как метагеномика. Его основной принцип состоит в том, что объектом исследования служат не отдельные микроскопические организмы, а их сообщества (микробиомы), при этом из отобранной пробы выделяется суммарная ДНК, секвенирование которой дает представление о наборе генов всех микроорганизмов, то есть о микробиоме в целом, а не только тех организмах, которых удается культивировать в лабораторных условиях или идентифицировать микроскопическими или микробиологическими методами.

Расширив таким образом поле зрения мировой науки, метагеномика позволила описать значительное количество новых групп на всех таксономических уровнях. Характерным примером выступает недавно открытая группа CPR (Candidate phyla radiation), ни один вид в которой на данный момент не изолирован в культуре. Согласно геномным данным, её представители отличаются набором рибосомных белков, отсутствием ряда ключевых метаболических путей и наличием самосплайсирующихся интронов в генах 16S рРНК (Brown *et al.* 2015). Филогенетический анализ указывает на то, что эта группа является сестринской всем остальным бактериям, а уровень дивергенции в её пределах не уступает таковому среди бактерий, не говоря уже об эукариотах (Hug *et al.* 2016).

Существуют два основных вида метагеномных исследований. Первый, более простой, называется анализом ампликонов. При этом амплифицируется и секвенируется определённый таксономический маркер, универсальный для исследуемых видов. Как правило, в роли такого маркера выступает последовательность гена малой субъединицы рибосомной РНК, этот ген широко используется в филогенетике и для него доступны многочисленные референсные последовательности. Например, широко используемая в анализе ампликонов база данных SILVA (Quast *et al.* 2013), в релизе 128 содержит 645 151 уникальную рРНК. Прочтения, полученные при секвенировании выделенной из исследуемого образца ДНК, сравниваются с такими базами данных, что позволяет отнести их к тому или иному таксону разного систематического уровня и получить информацию о разнообразии микробиоты в исследуемой среде.

Вторая разновидность, известная как метагеномика методом «дробовика» (shotgun metagenomics), основывается на секвенировании всей имеющейся в пробе ДНК, а не конкретных локусов. При достаточном покрытии такой подход позволяет описать не только таксономический состав сообщества, но и присутствующие в геномах представителей сообщества гены функциональных или структурных белков, в том числе – вирусных (Paez-Espino *et al*. 2016). На основе метагеномных данных можно устанавливать метаболические взаимодействия в отдельных микробиомах, опираясь на базы данных ePGDBs (environmental pathway/genome databases) (Hanson et al 2014). В некоторых работах из метагеномных наборов прочтений удавалось выделить полные геномы отдельных видов (Iverson *et al.* 2012).

В последние годы ампликонный анализ нашёл себе применение и в исследованиях микробиоты различных сред в озере Байкал. Описаны подлёдные бактериальные сообщества, связанные с цветением диатомей (Bashenkhaeva *et al.* 2015) и бактерии фотического слоя в весенний период (Михайлов и др. 2015). Исследованы бактерии, обитающие на байкальских губках (Калюжная, Кривич, Ицкович 2012, Гладких и др. 2014). Наконец, изучены бактериальные сообщества донных осадков в районах выхода углеводородов (Bukin *et al.* 2016, Zemskaya *et al.* 2015).

Для обеспечения метагеномных исследований требуются значительные вычислительные ресурсы, а также участие квалифицированного биоинформатика в обработке и интерпретации данных. Используемое программное обеспечение анализа ампликонов включает в себя различные библиотеки модулей обработки последовательностей, например, Mothur (Schloss *et al.* 2009), USearch (Edgar 2010), статистические пакеты и среды разработки алгоритмов многомерного статистического анализа данных, например, R ([https://www.r-project.org](https://www.r-project.org/)). Для проведения исследований с использованием обработки и анализа метагеномных данных специалисту требуются навыки составления сценариев в командной оболочке операционной системы (Linux, Windows), запуска пакетов в распределенной вычислительной среде и кластерных вычислительных системах, а также программирования на языках общего назначения, как правило, R или Python.

Другая важная проблема - это организация централизованного хранилища данных и обеспечение удобного к ним регламентированного доступа пользователям. На данный момент сотрудниками ЛИН СО РАН проведено множество ампликонных исследований самых разных экотопов Байкала, данные по некоторым из них собирались в течение нескольких лет. Тем не менее, нет строгих правил хранения исходных и промежуточных данных, а также полученных результатов. Сравнение и интеграция данных разных исследований также осложняется их неоднородностью, возникающей по причине использования различного программного обеспечения. Создание системы, обеспечивающей хранение исходных данных, метаданных и результатов метагеномных исследований в унифицированном виде, позволит упростить интеграцию результатов разных исследований и сравнительный анализ.

Целью данного исследования является разработка математического и программного обеспечения для поддержки процессов секвенирования нового поколения организационными, информационными и вычислительными ресурсами.

**РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ**

**Организация метагеномного анализа как вычислительного процесса**

Анализ предметной области показал, что задачи, решаемые как в биоиформационной части метагеномного анализа, так и в NGS в целом, хорошо представляются в рамках парадигмы Больших Данных. На данный момент научным сообществом разработаны форматы данных для хранения метагеномной информации; алгоритмы и программные модули, включая параллельные и распределенные реализации на кластерных вычислительных системах, обеспечивающие различные этапы анализа данных.

Решение задач в рамках парадигмы Больших Данных требует от исследователя-предметника (биолога) квалификации профессионального программиста-биоинформатика, так как исследования каждой пробы требуют составления отдельной программы-сценария или исполнения сценария вручную, что существенно замедляет процесс получения конечного результата.

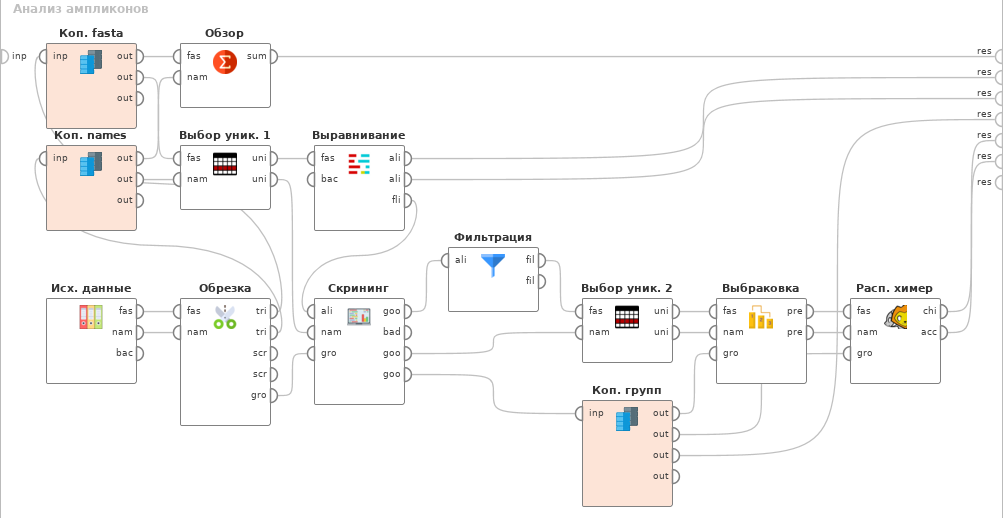
Организация исследований базируется на создании информационно-вычислительной среды, позволяющей проектировать и воспроизводить сценарии, подавая на вход данные различных форматов из различных источников, например, файлов, баз данных, серверов метагеномной информации. Среда должна обеспечивать также организованное облачное хранение промежуточных данных и результатов. Разработке такой среды поддержки научных исследований посвящен совместный проект ЛИН СО РАН и ИДСТУ СО РАН. В проекте необходимо решить несколько основных задач.

1. Исследование предметной области и ее декомпозиция на классы задач (функциональное моделирование). В результате такого моделирования выделяются классы задач и представляются в виде программных модулей. Из модулей формируются сценарии решения задач, сетевые графы модулей, связанные друг с другом передачей данных.
2. Разработка метаописаний модулей и структур входных и выходных данных. На этом этапе необходимо решить задачи интеграции с внешними информационными и вычислительными ресурсами. В связи с этим становится критически необходимым использовать стандарты и стандартные средства моделирования данных, в частности, онтологии.
3. Декомпозиция структур входных/выходных данных и реализация подсистем их преобразования, накопления, хранения и обеспечения эффективного (по критериям времени и вычислительной сложности) к ним доступа.
4. Для модулей, исходные коды которых недоступны из-за отсутствия исходного кода или лицензионных ограничений, строятся виртуальные среды и программные интерфейсы.
5. Разработка специализированного интерфейса пользователя для высокоуровневого управления сценарными расчетами. На этом этапе для обеспечения гибкости управления вычислительным процессом необходимо разработать интерфейс визуальной настройки сценария расчета.
6. Разработка подсистем визуализации и интерпретации полученных расчетов, включая модули интерпретации процесса расчета.

**Представление сценариев анализа метагеномных данных**

Популярным подходом к представлению вычислительного процесса, который строится в виде комбинации исполняющих модулей, выступает *программирование потоков данных* (*dataflow programming*) (Johson *et al*. 2004). Модули получают на вход некоторые данные, обрабатывают их и выдают на выход. Результаты вычислений передаются на вход других модулей. Подход развивается с 1970-х годов.

Подход удобен тем, что модули и сценарий их исполнения удобно задавать при помощи визуального редактора. Это позволяет создавать среды автоматизации решения задач пользователями специалистами-предметниками. Пользователь выбирает нужный модуль из набора модулей, добавляет его в рабочую область. Затем необходимо сцепить входы и выходы модулей (порты), обеспечив таким образом передачу информации. В результате должен получиться сетевой граф, не содержащий циклов и имеющий, как правило, один вход и один выход. Вход графа подаются данные, требующие обработки, на выход графа выдается результат. На рисунке 1 представлен пример использования разрабатываемой в проекте системы управления вычислительным процессом анализа ампликонов.



***Рис. 1.*** Сетевой граф процесса анализа ампликонов (начало).

Представленный сценарий сформирован при помощи программного комплекса Rapidminer Studio (https://rapidminer.com/), дополненного разработанными нами модулями для представления этапов анализа ампликонов. В сценарий входят следующие операции:

* формирование исследовательского проекта в виде папки (директории), содержащей данные секвенирования (модуль “Исх. данные”);
* обрезка данных (прочтений) (модуль “Обрезка”);
* модуль “Обзор” используется для визуального анализа качества результатов предыдущих шагов;
* сокращение объема входных данных за счет удаления незначащей информации, например, дублирующих последовательностей (модули “Выбор уник.”);
* выравнивание последовательностей на референсные базы данных (модуль “Выравнивание”);
* фильтрация последовательностей по заданным критериям (модуль “Скрининг”);
* удаление колонок выравнивания по заданным критериям, например, пустых колонок (модуль “Фильтрация”);
* удаление последовательностей, содержащих ошибки секвенирования (модуль “Выбраковка”);
* обнаружение химер (модуль “Расп. химер”, распознавание химер) и т.п.

В схеме представлены также сервисные модули RapidMiner, которые необходимы для распространения однотипной информации между модулями (“Коп. Групп”, копирование групп). Наличие таких модулей - это особенность программной системы, предполагающей, что в общем случае модули вносят изменения в обрабатываемую информацию без её копирования.

Каждый модуль получает на вход название файла и в результате работы создает новый файл. Функционирование модуля зависит от параметров, задаваемых пользователем при помощи пользовательского интерфейса. Результаты работы сценария передаются на специальные выходные порты и отображаются системой Rapidminer Studio в удобном пользователю виде. Система поддерживает возможность представления сценария в виде нового блока со своими портами ввода и вывода, а также возможность облачного хранения и исполнения сценариев, что позволяет создавать распределенные вычислительные среды. Богатый набор функций RapidMiner, а также различные сервисы, предоставляемые разработчиками стали основной причиной выбора этой системы в качестве среды разработки информационно-вычислительных ресурсов проекта.

**Хранилище метагеномной информации**

Изучение мирового опыта организации научных исследований в области Data Science (Наука, базирующася на анализе данных) показало, что использование облачных технологий является необходимой основой организации взаимодействия исследователей. В такую среду для обеспечения эффективного доступа пользователей и вычислительных процессов к данным исследований необходимо включать специализированное хранилище данных.

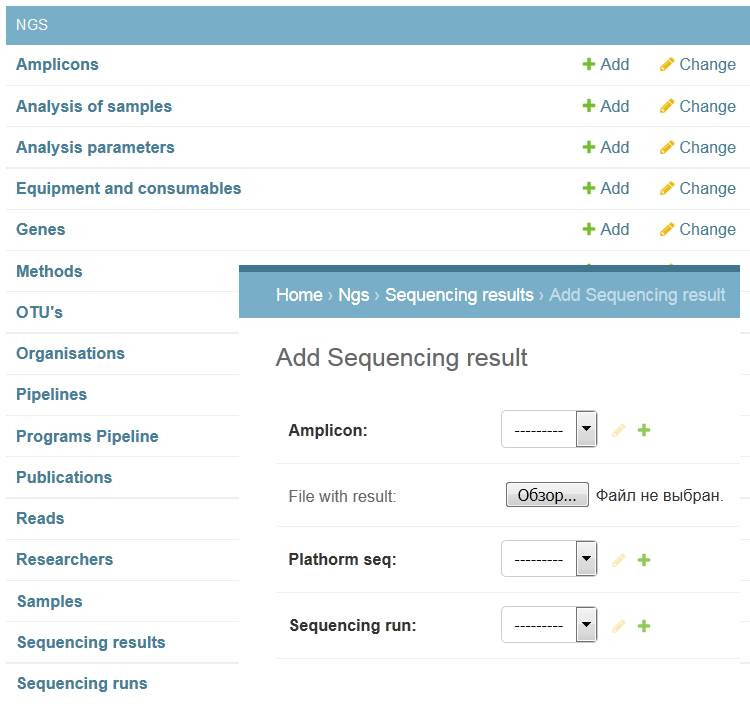
База данных для хранения данных исследований микробиома на основе метагеномного анализа (Рис. 2.) предусматривает хранение данных всего цикла исследования микробиома, начиная от отбора проб и заканчивая публикацией научных результатов. Схема, изображенная на рисунке 2 представляет собой модель Чена структуры базы данных. В схеме прямоугольниками обозначены хранимые в базе данных сущности, а ребрами - характеристики отношений между этими сущностями: один-ко-многим (подписываются как “1..n”), и многие-ко-многим (подписываются как “n..m”).

Схема охватывает информацию о сборе проб, анализе физико-химических и биологических параметров этих проб, результатов секвенирования, применяемом оборудовании и программном обеспечении, таксономических базах, методиках анализа собранных материалов, публикациях результатов исследований, а также исследователях, принимающих участие в получении результатов. Также она позволяет хранить используемые цепочки обработки и анализа метагеномных данных, включая информацию о программных инструментах, командах и конфигурационных файлов. Последняя функция позволяет сохранять состояние вычислительного процесса и перезапускать его с заданной точки.



***Рис. 2.***Обобщенное представление схемы данных поддержки исследований микробиома методами метагеномного анализа.

Модель реализована средствами инструментальной платформы разработки программного обеспечения Django ([https://www.djangoproject.com](https://www.djangoproject.com/)). Платформа Django обеспечивает автоматическое создание таблиц, представляющих отношения многие-ко-многим и настраиваемый интерфейс административной панели, позволяющий на этапе разработки проводить тестирование разработанной модели. Этими же средствами осуществляется реализация веб-сайта проекта (рис. 3).



***Рис. 3.***Снимок экрана пользовательского веб-интерфейса доступа к данным разработанной модели.

Облачное хранение информации и специализированное хранилище метагеномной информации позволяют создавать он-лайн сервисы для совместной обработки данных секвенирования из разных исследований специализированным программным обеспечением, осуществлять публикацию информации в Интернет. Для этого необходимо осуществить:

* разработку программного интерфейса доступа к данным;
* наполнение базы данных информацией, собранной и обработанной в результате исследований микробиома озера Байкал в 2009-2015 годах;
* автоматизацию процессов планирования и исполнения сценариев обработки и анализа метагеномных данных в распределенной вычислительной среде.

**Заключение**

В статье представлены современные проблемы организации распределенной программной среды для создания организационных, информационных и вычислительных ресурсов проведения научных микробиологических исследований на основе метагеномного анализа. Проведена обобщенная модель предметной области системного уровня и выделены требования к разрабатываемой среде, задачи, требующие решения, построена и частично реализована модель вычислительного процесса анализа ампликонов. Аспект информационной поддержки вычислительного процесса представлен задачей создания облачного хранилища вычислительных процессов (сценариев исследований), а также базы данных, предназначенной для хранения данных (исходных, промежуточных) и результатов. На основе базы данных решается задача построения информационного портала для обработки метагеномных данных и представления результатов научному сообществу.

**Благодарности.** Работа выполняется в рамках проекта Иркутского научного центра СО РАН № 4.1.2 «Применение методов NGS-BD (Next Generation Sequencing – Big Data) – для решения вопросов экологии». Результаты получены при активном использовании сетевой инфраструктуры Телекоммуникационного центра коллективного пользования “Интегрированная информационно-вычислительная сеть Иркутского научно-образовательного комплекса” (ЦКП ИИВС ИРНОК) (http://net.icc.ru).

# **Литература**

Гладких А.С., Калюжная О.В., Белых О.И., Ан Т.С., Парфёнова В.В. Анализ бактериального сообщества двух эндемичных видов губок из озера Байкал // Микробиология. — 2014. — Т. 83, № 6. — С. 682-693.

Михайлов И.С., Захарова Ю.Р., Галачьянц Ю.П. и др. Об однородности таксономического состава бактериальных сообществ фотического слоя трёх котловин озера Байкал, различающихся по составу и обилию весеннего фитопланктона // Доклады Академии Наук. — 2015. — Т. 465, № 5. — С. 620-626.

Bashenkhaeva M.V., Zakharova Y.R. Petrova D.P. Sub-Ice microalgal and bacterial communitie in freshwater Lake Baikal, Russia // Environmental Microbiology. — Vol. 70, No. 3. — P. 751-765.

Brown C.T., Hug L.A., Thomas B.C. *et al.* Unusual biology across the group comprising more than 15% of domain Bacteria // Nature. — 2015. — Vol. 523. — P. 208-211.

Bukin S.V., Pavlova O.N., Manakov A.Y. *et al.* The ability of microbial community of Lake Baikal bottom sediments associated with gas discharge to carry out the transformation of organic matter under thermobaric conditions // Frontiers in microbiology. — 2016. — Vol. 7. — P. 690.

Edgar R.C. Search and clustering orders of magnitude faster than BLAST // Bioinformatics. — 2010. — Vol. 26, No. 19. — P. 2460-2461.

Hanson N.W., Konwar K.M., Wu S.J., Hallam S.J. Introduction to the analysis of environmental sequence information using metapathways // Comp. Meth. Next Gener. Sequenc. Data Analysis. – 2016. – P. 25–56.

Hug L.A., Brett J.B, Anantharaman K. *et al.* A new view of the Tree of Life // Nature Microbiology. — 2016. — Vol. 1. — P. 16048.

Iverson V., Morris P.M., Frazar C.D. Untangling genomes from metagenomes: Revealing an uncultured class of marine Euryarchaeota // Science. — 2012. — Vol. 335. — P. 587-590.

Johnston W.M., Hanna J.R.P., Millar R.J. [Advances in Dataflow Programming Languages //](http://www.cs.ucf.edu/~dcm/Teaching/COT4810-Spring2011/Literature/DataFlowProgrammingLanguages.pdf) ACM Computing Surveys. – 2004. – Vol. 36. – P. 1–34.

Paez-Espino D., Eloe-Fadrosh E.A., Pavlopoulos G.A. et al. Uncovering earth`s virome // Nature. – 2016. – Vol. 536. – P. 425–430.

Quast C., Pruesse E., Yilmaz P. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tool // Nucleic Acids Research. — 2013. — Vol. 41. — P. 590-596.

Schloss P.D. *et al.* Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities // Applied and Environmental Microbiology. — 2009. — Vol. 75 (No. 23). — P. 7537-7541.

Zemskaya T.I, Lomakina A.V., Mamaeva E.V. *et al.* Bacterial communities in sediments of Lake Baikal from areas with oil and gas discharge //Aquatic Microbial Ecology. — 2015. — Vol. 75. — P. 95-109.

**Computational infrastucture for supporting the metagenomic analysis**

Morozov A.A.1, 2\*, Shigarov A.O.2, 3, Malkov F.S. 2, 3, Paskal K.K. 2, 3,

Cherkashin E.A. 2, 3, Likhoshway Ye.V.1

1 Лимнологический институт СО РАН, г. Иркутск

2Иркутский научный центр СО РАН, г. Иркутск

3Институт динамики систем и теории управления им. В.М. Матросова СО РАН, г. Иркутск

morozov@lin.irk.ru

Metagenomic analysis, while allowing to describe the microbial communities with an unprecedented precision, requires substantial computational resources and a competent bioinformatician. These requirements have slowed down its wide adoption. We have started the development of the nonspecialist-targeted system for the storage and analysis of the metagenomic data which will give the researchers a convenient analysis interface and facilitate the unified storage of raw data and analysis results. This paper presents the dataflow programming system and the schema of the SQL database used for metadata storage.

Keywords: bioinformatics, metagenomic analysis, Big Data