ФЕДЕРАЛЬНОЕ АГЕНТСТВО НАУЧНЫХ ОРГАНИЗАЦИЙ

|  |  |
| --- | --- |
| ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ УЧРЕЖДЕНИЕ НАУКИ  ИРКУТСКИЙ НАУЧНЫЙ ЦЕНТР СИБИРСКОГО ОТДЕЛЕНИЯ  РОССИЙСКОЙ АКАДЕМИИ НАУК (ИНЦ СО РАН) | |
| УДК 574.5+004.5  № госрегистрации: АААА-А17-117041250054-8  Инв. № 2017-01 | УТВЕРЖДАЮ  Директор ИНЦ СО РАН д.м.н.  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ К.А. Апарцин  «\_\_\_\_\_ » декабря 2018 г. |
| ОТЧЕТ  о научно-исследовательской работе  по теме  ФУНДАМЕНТАЛЬНЫЕ ИССЛЕДОВАНИЯ И ПРОРЫВНЫЕ ТЕХНОЛОГИИ  КАК ОСНОВА ОПЕРЕЖАЮЩЕГО РАЗВИТИЯ БАЙКАЛЬСКОГО РЕГИОНА  И ЕГО МЕЖРЕГИОНАЛЬНЫХ СВЯЗЕЙ  (0341-2016-0001)  (промежуточный)  Направление 4 - Научное обоснование охраны озера Байкал и социально-экономического развития Байкальской природной территории  Часть 4.2  Применение методов NGS-BD (Next Generation Sequencing – Big Data) для решения вопросов экологии (часть 2)  Руководитель темы  академик РАН И.В. Бычков | |

Иркутск 2018

СПИСОК ИСПОЛНИТЕЛЕЙ

Научные руководители части 4.2:

Лихошвай Е.В., д.б.н., проф.,

зав. отд. Ультраструктуры клетки ЛИН СО РАН Разделы 1-9

Земская Т.И., д.б.н.,

зав. лаб. Микробиологии углеводородов ЛИН СО РАН Разделы 1,4

Галачьянц Юрий Павлович \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Разделы 2, 3, 6

к.б.н., с.н.с. подпись, дата

Петрова Дарья Петровна \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Разделы 1, 3

к.б.н., н.с. подпись, дата

Морозов Алексей Анатольевич \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Раздел 5

гл. спец. биоинф. подпись, дата

Павлова Ольга Николаевна \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Раздел 4

к.б.н., с.н.с. подпись, дата

Ломакина Анна Владимировна \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Раздел 1

к.б.н., с.н.с. подпись, дата

Башенхаева Мария Викторовна \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Раздел 3

к.б.н., с.н.с. подпись, дата

Букин Юрий Сергеевич \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Разделы 2, 3, 9

к.б.н., с.н.с. подпись, дата

Черкашин Евгений Александрович, \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Разделы 7, 8, 9

к.т.н., с.н.с., доц. подпись, дата

Шигаров Алексей Олегович \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Разделы 5, 7

к.т.н., с.н.с. подпись, дата

Малков Фёдор Сергеевич \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Раздел 5

м.н.с. подпись, дата

Нормоконтролер Чагина Ольга

Мстиславовна к.ф.-м.н. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

подпись, дата

РЕФЕРАТ (1стр!!!!)

Отчет 56 с., 9 разделов, рис., 4 таблицы, 34 источника, 3 прил.

Применение методов NGS-BD (Next Generation Sequencing – Big Data)  
для решения вопросов экологии

*Перечень ключевых слов*: анализ больших объемов данных, информационные технологии и системы, распределенные информационно-вычислительные ресурсы, экоинформатика, базы данных и знаний, водные экосистемы, секвенирование нового поколения, микробиом, корреляционные сети, динамическая система, порождающее программирование, моделирование вычислительных процессов

*Объектами исследования и разработки* являются микробиом озера Байкал и его физико-химические характеристики, процессы секвенирования нового поколения (NGS).

*Цель исследования* – реализация системного подхода к исследованию озера Байкал, основывающегося на анализе и интерпретации данных NGS и реализующего современные мировые стандарты и инструментальные средства.

*Методы исследования*: исследование микробиома основывается на получении первичного материала (проб) их предварительной обработке согласно выбранной методике с целью выделения ДНК, проведению секвенирования на специализированном оборудовании, анализе полученных данных на основе существующих программ-сценариев NGS. Затем исследуются и интерпретируются корреляционные зависимости между характеристиками микробиома. Результатом интерпретации выступает идентификация соответствующей динамической модели.

Процесс проведения секвенирования нового поколения подвергается системному анализу информационных потоков, процедур обработки данных и взаимодействий с внешними программными системам; разрабатываются графовые модели этих взаимодействий и структуры данных для облачного хранения исходных данных и результатов исследований.

*Результаты исследования*: внедрена методика подготовки проб для секвенирования ДНК, основанная на ферментативной обработке клеток; разработана консервативная процедура тримминга исходных прочтений, позволяющая оказывать меньшее влияние на таксономическое разнообразие данных микробиома; спроектированы структуры данных и визуальные модели для поддержки научных исследований в NGS, интеграции данных с мировыми информационными ресурсами; разработана методика идентификации динамической модели физико-химических процессов альго-бактериальных взаимодействий.

*Рекомендации по внедрению* результатов научных исследований и разработок будут сделаны после завершения работ.

*Область применения* результатов исследований относится к оценке экологического состояния озер.

*Цель проекта* – характеристика микробиома озера Байкал на основе методов секвенирования нового поколения (NGS) и результатов обработки полученной информации аналитическими технологиями Больших Данных (Big Data) для решения задач экологии.

*Задачи, обеспечивающие достижение цели*:

Создание программно-аппаратной Инфраструктуры Обработки Больших Данных (ИОБД) для архивирования, курирования, анализа и визуализации сверхбольших массивов информации, получаемой в результате высокопроизводительного секвенирования при исследовании микробиома Байкала.

Создание коллекции суммарной ДНК из водной толщи, донных осадков и ДНК из отдельных видов для последующего использования.

Выявление структуры и построение корреляционных взаимосвязей в метасообществах оз. Байкал по данным анализа фрагментов генов 16S и 18S рРНК, определенных функциональных генов и реконструкций метагеномов в продуктивных зонах, на границах раздела фаз и по градиентам физико-химических факторов окружающей среды.

Выделение и описание новых видов бактерий из разных экотопов озера Байкал, перспективных для практического применения, определение их метаболического потенциала на основе анализа их геномов и предсказанных протеомов.

Экспериментальная апробация биоинформатических методов анализа прокариотических и эукариотических сообществ в среде ИОБД. Опытное построение и выполнение соответствующих технологических цепочек аналитики данных средствами ИОБД. Создание унифицированной метагеномной базы данных как основы платформы «Микробиом Байкала» для решения вопросов экологии.

Разработка распределенной динамической математической модели взаимосвязи физико-химических параметров среды и таксономического состава микробиома.

* 1. СОДЕРЖАНИЕ

СОДЕРЖАНИЕ 3

ВВЕДЕНИЕ 5

1 Управление данными информационно-аналитической системы поддержки исследования микробиома озера Байкал 8

1.1 Материалы и методы 8

1.2 Результаты с таблицами и картинками 9

Рисунок 1.2 – Упрощенное представление схемы данных исследования  
микробиома озера Байкал 9

Рисунок 1.3 – Программа подготовки финального списка таблицы встречаемости OTU 10

Рисунок 1.4 – Программа подготовки списка группы таблицы встречаемости OTU 11

Рисунок 1.5 – Программа запроса уникальных проб из списка проб 11

Рисунок 1.6 – Программа поиска совпадений заданного ампликона 11

Рисунок 1.7 – Блок-схема алгоритма генерации таблицы встречаемости OTU 12

Рисунок 1.8 – Фрагмент таблицы встречаемости OTU, полученной в результате запуска реализованного алгоритма 12

Рисунок 1.9 – Снимок экрана тестирования программного интерфейса (RESTful) доступа к данным исследования микробиома озера Байкал 13

Рисунок 1.10 – Пример с фрагментом ответа в формате JSON с результатом выполнения запроса получения всех доступных URL ссылок интерфейса 14

Рисунок 1.11 – Пример с фрагментом ответа в формате JSON с результатом выполнения запроса получения всех объектов класса Researcher. 14

Рисунок 1.12 – Пример с фрагментом ответа в формате JSON с результатом выполнения записи нового объекта класса Researcher 14

1.1 Краткое обсуждение 14

1.2 Заключение по своей части 15

1.3 Список опубликованных работ с указанием базы, в которой индексируется издание 15

2 Автоматизация 16

Рисунок 2.1 – Снимок экрана тестирования программного интерфейса (RESTful) доступа к данным исследования микробиома озера Байкал 16

ЗАКЛЮЧЕНИЕ 17

СПИСОК ИПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 19

Приложение А  
 Публикации по теме проекта за 2017 год 24

Приложение Б  
Информация о количестве статей в виде таблицы 26

ПРИЛОЖЕНИЕ В   
Этапы выполнения проекта 27

* 1. ВВЕДЕНИЕ

Основой структурно-функциональной организации водных экосистем являются первичные продуценты и прокариоты, они определяют биогеохимические циклы основных элементов и круговорот органических веществ. В настоящее время представление о полном таксономическом разнообразии одноклеточных эукариот, бактерий и архей, генетической структуры сообществ и их метаболического потенциала стало возможным получать благодаря внедрению методов секвенирования нового поколения и биоинформатических методов анализа больших объемов информации о структурах генов («больших данных»). Развитие программного обеспечения на основе методов многомерного анализа данных и теории графов позволяет установить корреляционные связи между отдельными выявленными таксонами про- и эукариот, между определенным таксоном и изменяющимися факторами окружающей среды. Этот системный подход начали внедрять в изучение морских водных экосистем с 2011 г. [1,2] и использовать в выполнении нескольких международных проектов (TARA Oceans, Global Ocean Survey, Hawaiian Ocean Time Series, Bermudan Ocean Time Series, the Long Term Ecological Research sites, NEON и расширяющийся проект the Earth Microbiome Project). Для определения структурно-функциональной организации оз. Байкала на современном уровне предлагается проведение исследований по аналогичному принципу.

К настоящему времени для озера Байкал получено и проанализировано большое количество библиотек генов 16S и 18S рРНК, а также некоторых функциональных генов, позволяющих охарактеризовать разнообразие прокариотических и эукариотических сообществ и их способность осуществлять определенные геохимические и метаболические процессы в водной толще и донных осадках озера Байкал. Исследование микробиома озера Байкал, выполняемое с помощью высокопроизводительного секвенирования нового поколения, приводит к получению сверхбольших массивов данных (Big Data). Они характеризуются: (1) большим объемом – в результате секвенирования материала даже одной пробы объемом 1 мл воды может генерироваться порядка 50 Гб информации; (2) высокой скоростью их накопления, например, ежегодный сезонный мониторинг микробиома Байкала требует обработки до нескольких сотен таких проб; и (3) сложностью многообразия исходных данных – каждая проба несет в себе информацию о нескольких тысячах различных видов микроорганизмов (бактерий, архей, вирусов, грибов, фито- и зоопланктона). На современном этапе развития технологий биоинформатики исследование микробиома оз. Байкал требует создания программно-аппаратной инфраструктуры обработки и аналитики Больших Данных для архивирования, курирования, анализа и визуализации получаемых сверхбольших массивов информации.

В прикладном аспекте проект направлен на исследование ферментных систем байкальских психрофильных микроорганизмов, обеспечивающих деструкцию или продуцирование разных соединений. Предполагается построить динамические модели сред стабильного функционирования таких ферментных систем.

**Одно предложение в общее резюме = основное достижение**

Выполнено прототипирование информационно-аналитической системы поддержки исследования микробиома озера Байкал, включая усовершенствованную схему представления и хранения исследовательских метагеномных данных и OTU (Operational Taxonomic Unit) результатов, алгоритм разбора и загрузки OTU результатов из формата пакета программного обеспечения обработки данных высокопроизводительного секвенирования MOTHUR (https://www.mothur.org) в базу метагеномных данных, веб-ориентированный RESTful интерфейс прикладного программирования для доступа к данным исследования микробиома на основе архитектуры взаимодействия компонентов распределённого приложения в сети.

# Управление данными информационно-аналитической системы поддержки исследования микробиома озера Байкал

Выполнено проектирование и прототипирование информационно-аналитической системы поддержки исследования микробиома озера Байкал. Усовершенствована схема представления и хранения исследовательских метагеномных данных и OTU результатов. Реализован процесс автоматического разбора и загрузки OTU-данных, полученных в программе обработки данных высокопроизводительного секвенирования MOTHUR в базу метагеномных данных. Разработан веб-ориентированный RESTful-интерфейс прикладного программирования для доступа к данным исследования микробиома на основе архитектуры взаимодействия компонентов распределённого приложения в сети. Интерфейс обеспечивает чтение и запись данных для всех объектов текущей модели информационно-аналитической системы. Развёрнута инфраструктура проекта, включая систему совместной разработки программного обеспечения (доступна по адресу https://git.icc.ru), систему управления проектом (доступна адресу https://redmine.icc.ru).

## Материалы и методы

Проектирование и реализация объектной модели информационно-аналитической системы поддержки исследования микробиома озера Байкал выполнена на платформе разработки программного обеспечения веб-приложений DJANGO с использованием шаблона проектирования MVC (Model-View-Controller) для разделения данных приложения, пользовательского интерфейса и управляющей логики. Проектирование усовершенствованной схемы данных осуществлялось в соответствии спостановкой задачи и предоставленных исследовательских данных ЛИН СО РАН. На основе разработанной схемы подготовлено метаописание таблиц базы данных. Сформированные классы объектной модели в терминах DJANGO позволили сгенерировать SQL-скрипты (Structured Query Language) создания структуры базы данных в формате СУБД (системы управления базами данных) PostgreSQL.

Для разработки программного интерфейса созданы управляющие скрипты: (1) сериализации (serializing) объектов модели разрабатываемой информационно-аналитической системы в формате JSON (JavaScript Object Notation); (2) создания представлений для каждого сериализатора класса этой модели; (3) подключения списка URL-ссылок доступных запросов интерфейса к платформе разработки Django; (4) настройки безопасности доступа к данным с поддержкой пользовательской авторизации.

## Результаты с таблицами и картинками

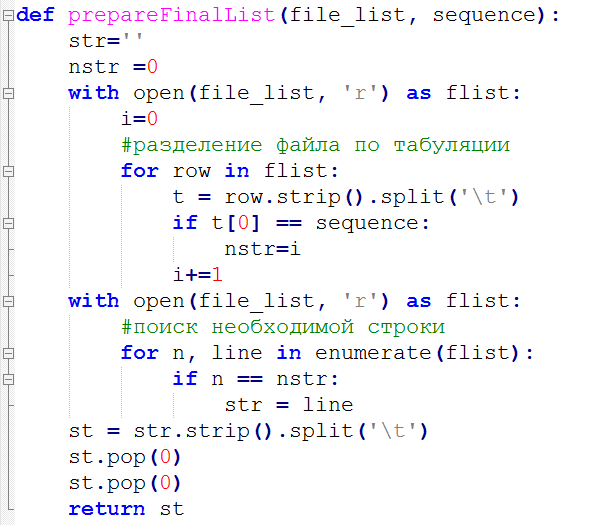
Усовершенствована схема данных исследования микробиома озера Байкал. На рисунке 1 представлена её упрощенная модель. Выполнено разделение на следующие модули:

* Модуль RPM обеспечивает общее описание проводимых исследовательских работ, включая данные по их исполнителям, публикациям, методикам, оборудованию и расходным материалам.
* Модуль WaterLIN служит для представления специфических данных исследования микробиома озера Байкал, включая описания отбираемых проб, измеримые параметры, результаты анализа проб, изучаемые ампликоны, полученные прочтения, результаты секвенирования.
* Модуль NGS организует данные, связанные с проведением вычислений по высокопроизводительному секвенированию ДНК/РНК последовательностей исследуемых проб воды и донных осадков, включая информацию по операционным таксономическим единицам, идентификаторам таксонов, используемым таксономическим базам данных, рабочим процессам (вычислительным цепочкам) обработки последовательностей и анализа результатов, списку программ, задействованных в этих рабочих процессах.



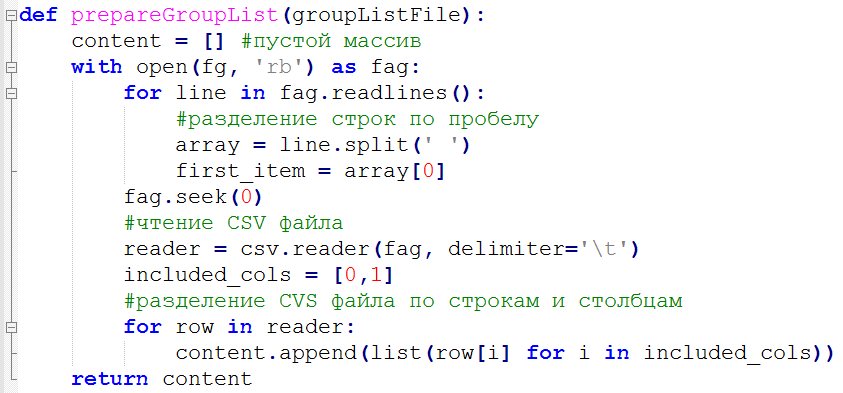
* + 1. Упрощенное представление схемы данных исследования  
       микробиома озера Байкал

Разработан алгоритм генерации таблицы встречаемости OTU по выходным данным формата MOTHUR. На вход алгоритм получает MOTHUR файлы final.an.list и final.groups. Исследователь выбирает необходимую частоту встречаемости, для которой необходимо посчитать OTU. На выходе алгоритм генерирует файл с необходимой таблицей встречаемости. Ниже представлены блок-схемы основных функций алгоритма (см. рисунки. 2-6). На рисунке 2 представлен алгоритма подготовки финального списка, которая из файла с множеством результатов частот встречаемости выделяет нужную строку с заданной исследователем частотой встречаемости.



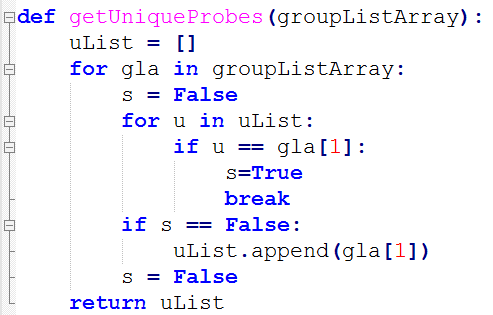
* + 1. Программа подготовки финального списка таблицы встречаемости OTU

На рисунке 3 показан алгоритм подготовки списка по файлу final.groups, который разделяет файл на строки и столбцы для последующего удобного поиска по файлу.



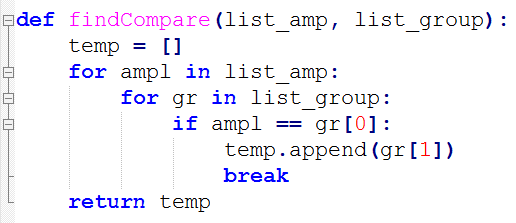
* + 1. Программа подготовки списка группы таблицы встречаемости OTU

Рисунок 4 демонстрирует функцию получения уникальной пробы, анализирует массив данных, полученных после подготовки списка по файлу final.groups и выдает на выходе список проб, состоящий из проб по содержимому файла final.groups.

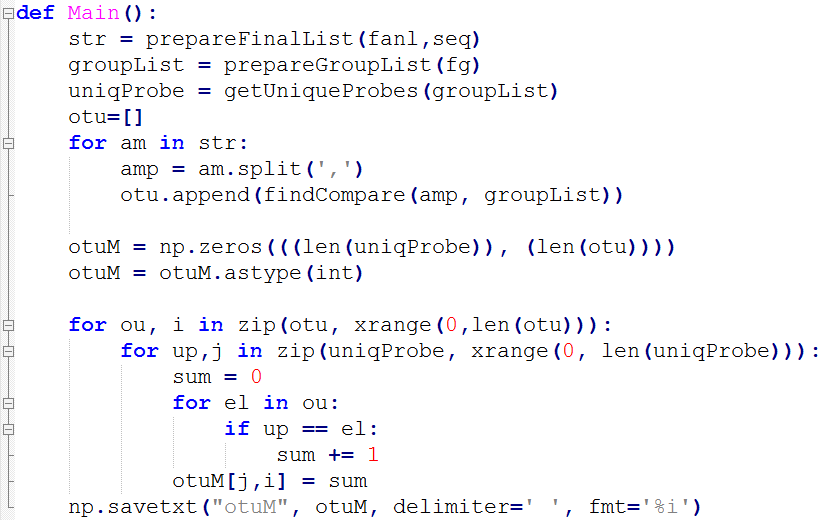


* + 1. Программа запроса уникальных проб из списка проб

На рисунке 5 представлен алгоритм поиска совпадений для заданного ампликона по данным, полученным из файла final.groups.

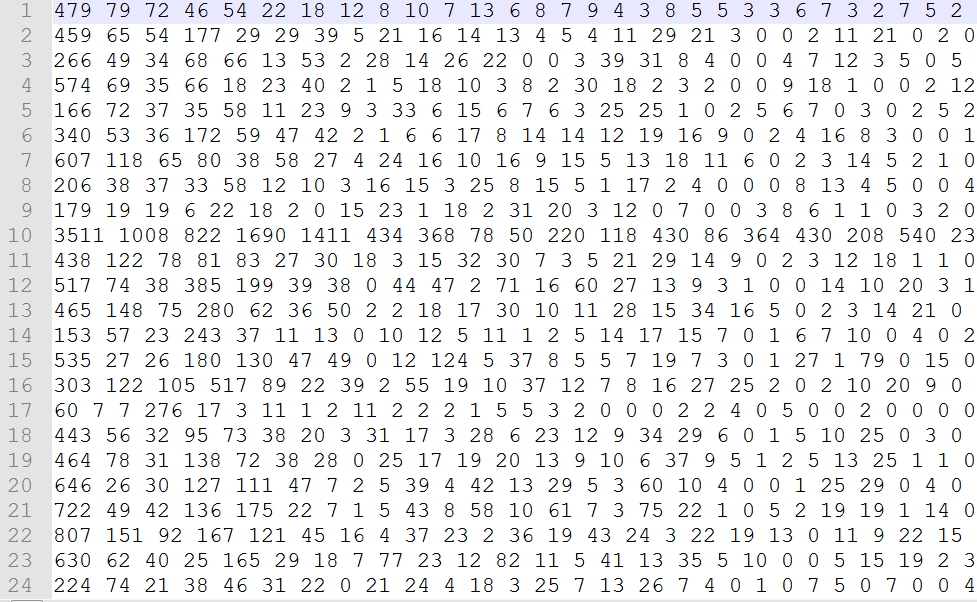


* + 1. Программа поиска совпадений заданного ампликона



* + 1. Блок-схема алгоритма генерации таблицы встречаемости OTU

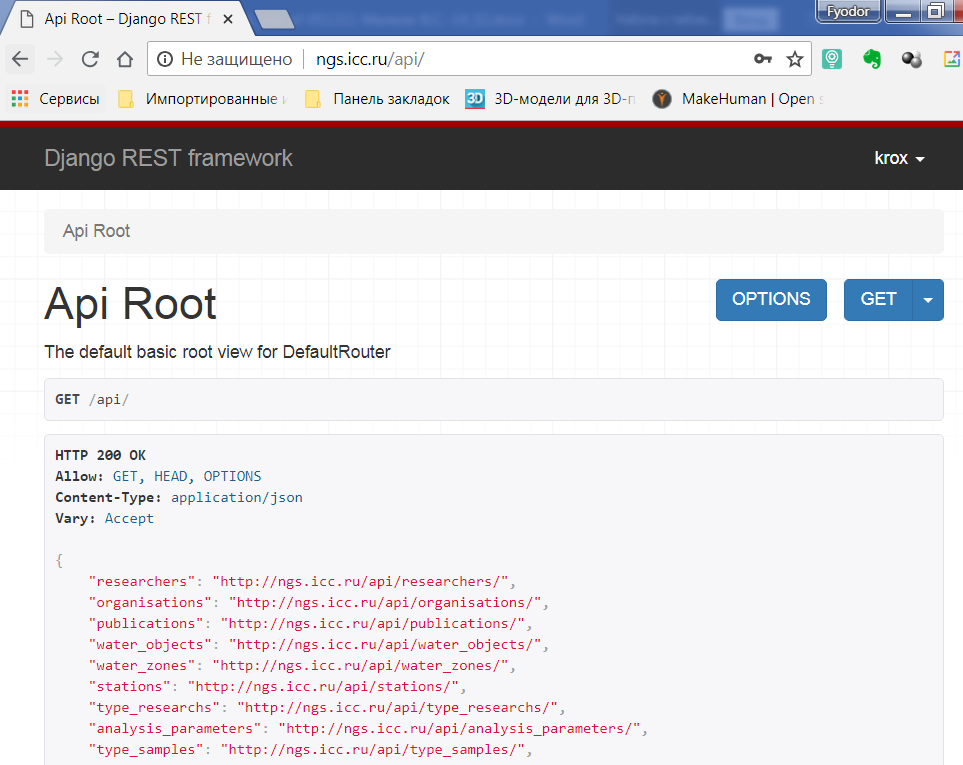
На рисунке 6 представлен алгоритм генерации таблицы встречаемости OTU, задающий рабочий процесс вызова функций, представленных на рисунках 2-5. Алгоритм считает количество ампликонов для каждой пробы, происходит генерация двумерной матрицы и её наполнение статистическими подсчётами, а также сериализацию результатов в файл.



* + 1. Фрагмент таблицы встречаемости OTU, полученной в результате запуска реализованного алгоритма

Спроектированный алгоритм реализован на языке программирования PYTHON. Фрагмент таблицы встречаемости OTU, полученной в результате запуска этой реализованного, показан на рисунке 7.

Спроектирован и реализован программный веб-ориентированный интерфейс доступа к данным исследования микробиома озера Байкал на основе архитектуры взаимодействия компонентов распределенного приложения REST (Representational State Transfer) API. Этот интерфейс обеспечивает чтение и запись данных для всех объектов модели разрабатываемой информационно-аналитической системы. Тестовая версия интерфейса (см. рисунок 8) доступна по адресу https://ngs.icc.ru/api.



* + 1. Снимок экрана тестирования программного интерфейса (RESTful) доступа к данным исследования микробиома озера Байкал

Интерфейс определяет, что для отправки HTTP-запроса необходимо указать URL запроса, тип (POST или GET), а также данные для авторизации (имя пользователя и пароль). В ответе REST-интерфейс возвращает результат выполнения запроса в формате JSON. Пример отправки запроса для получения всех доступных URL для классов модели с использованием утилиты CURL выглядит следующим образом: curl -X GET --user user:password http://ngs.icc.ru/api. В результате будет получен следующий список всех доступных URL в формате JSON, как показано ниже:

{"researchers":"http://127.0.0.1:8000/api/researchers/",  
"organisations":"http://127.0.0.1:8000/api/organisations/",  
"publications":"http://127.0.0.1:8000/api/publications/",  
"water\_objects":"http://127.0.0.1:8000/api/water\_objects/",  
"water\_zones":"http://127.0.0.1:8000/api/water\_zones/",  
"stations":"http://127.0.0.1:8000/api/stations/","type…

* + 1. Пример с фрагментом ответа в формате JSON с результатом выполнения запроса получения всех доступных URL ссылок интерфейса

Следующий пример показывает выполнение GET запроса для получения всех объектов класса Researcher, представляющего данные исполнителей исследовательских работ: curl -X GET --user user:password http://ngs.icc.ru/api/researchers. В результате выполнения данного запроса будет получен список всех исследователей, как показано ниже.

[{"user":1,"fio":"Ivanov","organisation":1,"lab":"laboratory 1","contacts":"890256660777","date\_added":"2018-08-02T01:26:39.586000+08:00"},{"user":null,"fio":"Petrov","organisation":1,"lab":null,

"contacts":null,"date\_added":"2018-12-07T10:43:31.039000+08:00"},{"user":null,"fio":"Ivanov","organisation":1,"lab":null,

"contacts":"892435556677","date\_added":"2018-12-07T11:14:29.374000+08:00"}]

* + 1. Пример с фрагментом ответа в формате JSON с результатом выполнения запроса получения всех объектов класса Researcher.

Дополнительно к запросам чтения данных, интерфейс также предоставляет возможность изменения данных. Например, для записи данных необходимо отправить запрос POST и указать необходимые параметры объекта. Пример записи нового объекта класса Researcher: curl –X POST –d “fio=Ivanov&organisation=1&contacts=892435556677” --user user:password http://ngs.icc.ru/api/researchers, получаемым ответом, как показано ниже.

{"user":null,"fio":"Ivanov","organisation":1,"lab":null,"contacts":"892435556677", "date\_added":"2018-12-07T11:14:29.374000+08:00"}

* + 1. Пример с фрагментом ответа в формате JSON с результатом выполнения записи нового объекта класса Researcher

## Краткое обсуждение

При исследовании микробиома озера Байкал методами высокопроизводительного секвенирования порождаются большие массивы метагеномных данных. Жизненный цикл этих данных включает этапы сбора, обработки, анализа, публикации и архивирования этих. Микробиологи не всегда способны самостоятельно организовать управление жизненным циклом своих исследовательских данных. Планирование и управление данными требует навыков работы со специализированным программным обеспечением. Предполагается, что разрабатываемая информационно-аналитической системы поддержки исследования микробиома озера Байкал обеспечит микробиологов возможностью упрощенного управления метагеномными данными на протяжении всего их жизненного цикла. В частности, она позволяет микробиологам хранить и публиковать в открытом и защищенном доступе свои первичные и обработанные данные, результаты исследования, а также самостоятельно создавать и хранить собственные рабочие процессы обработки и анализа метагеномных данных при исследовании микробиома озера Байкал.

## Заключение по своей части

Разработаны средства управления данными (в т. ч. схема данных и объектная модель информационно-аналитической системы поддержки исследования микробиома озера Байкал, веб-ориентированный пользовательский и программный интерфейс доступа к исследовательским данным) с целью обеспечения курирования больших объемов данных высокопроизводительного секвенирования.

## Список опубликованных работ с указанием базы, в которой индексируется издание

1. Cherkashin E., Shigarov A., Malkov F., and Morozov A. An Instrumental Environment for Metagenomic Analysis. Proc. Int. Conf. on Information Technologies in the Research of Biodiversity. 2018. *(Принято в печать в издании, индексируемом Web of Science Core Collection)*

**Список процитированной литературы**

нет

# Визуальное моделирование проблемно-ориентированных пакетов

**Результат в одну строку**

Созданы средства автоматизации построения визуальных моделей потоков данных для прикладных пакетов биоинформатики.

**Краткий результат**

Разработана программная система анализа исходного кода прикладного пакета Mothur, создающая абстрактную модель модулей в виде графа отношений элементов структуры пакета. Создана и формализована система знаний, при помощи которой по результатам вышеуказанного анализа осуществляется порождение исходного кода Java-модулей для системы визуального проектирования вычислительных процессов Rapidminer. В результате в Rapidminer представлены 143 процедуры обработки данных NGS. Разработанная технология позволит на следующих этапах проекта также сгенерировать визуальное представление модулей вычислительных сред NGS, таких систем, как QIIME, QIIME2, Usearch и др., что позволит унифицировать модель вычислительной процедуры анализа ампликонов и сделать ее независимой от используемого прикладного пакета.

**Введение**

Используемое в научных исследованиях ЛИН СО РАН программное обеспечение анализа ампликонов включает в себя различные библиотеки модулей обработки последовательностей, например, Mothur, USearch, статистические пакеты и среды разработки алгоритмов многомерного статистического анализа данных, например, R (https://www.r-project.org). Для проведения исследований с использованием обработки и анализа метагеномных данных специалисту требуются навыки программирования (составления сценариев) в командной оболочке операционной системы (Linux, Windows), запуска пакетов в распределенной вычислительной среде и кластерных вычислительных системах, а также программирования на языках R или Python. Одним из способов расширения круга пользователей сложных прикладных пакетов – это представление структуры модулей прикладного пакета в виде визуальной модели, а также разработки средств визуального проектирования вычислительного процесса. Ы результате пользователи-предметники, не являющиеся специалистами в области программирования, смогут самостоятельно формировать вычислительную процедуру из визуальных блоков.

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

Подход удобен тем, что модули и сценарий их исполнения удобно задавать при помощи визуального редактора. Пользователь выбирает нужный модуль из существующего набора, 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. На вход графа подаются данные, требующие обработки, на выход графа выдается результат, который сохраняется в файл или визуализируется при помощи соответствующих модулей.

Каждый модуль прикладного пакета Mothur получает на вход название файла и в результате работы создает новый файл. Функционирование модуля зависит от параметров, задаваемых пользователем индивидуального для каждого модуля. Разработанные программные средства позволяют представлять вычислительный процесс Mothur в виде блоков системы Rapidminer Studio, реализующих вышеупомянутый модель потоков данных. Система поддерживает возможность представления сценария в виде нового комплекного 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.

**Материалы и методы**

На этапе 2018 года направлением развития автоматизации программирования сценариев метагеномного анализа являлась разработка системы интерпретации метаописаний операций программы Mothur в промежуточное представление, разработка процедуры трансформации данного представления в исходный код модулей системы Rapidminer и формирование библиотеки вычислительных модулей, которой могу пользоваться специалисты-предметники для решения биоинформатических задач.

Необходимость в автоматическом анализе исходного кода Mothur и синтезе соответствующих структур модулей Rapidminer определяется, во-первых, количеством модулей, которые необходимо представить в виде визуальных блоков, – 143, а, во-вторых, необходимостью поддерживать соответствие сгенерированных блоков новым версиям модулей Mothur, т.к. система находится в фаза активного развития. Кроме того, наличие такой модели позволяет на следующих этапах проекта разрабатывать процедуры автоматизации вычислительного процесса на кластерных вычислительных системах.

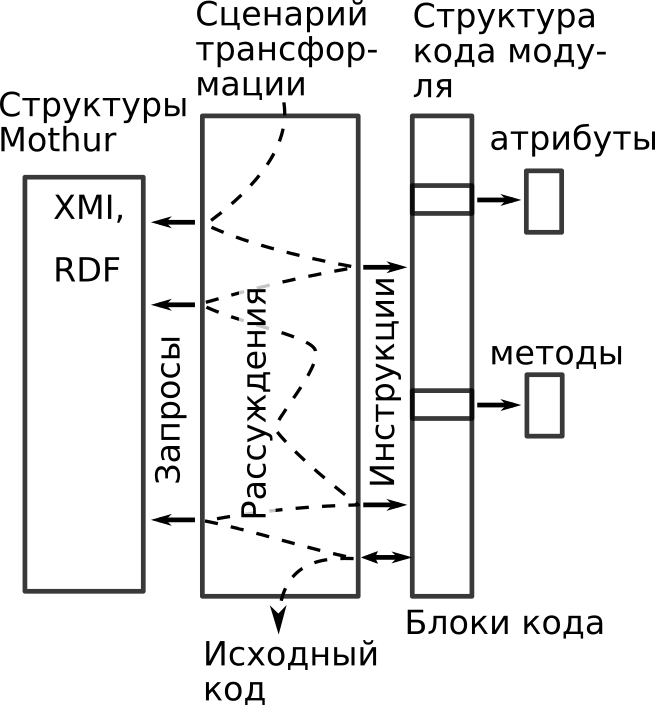
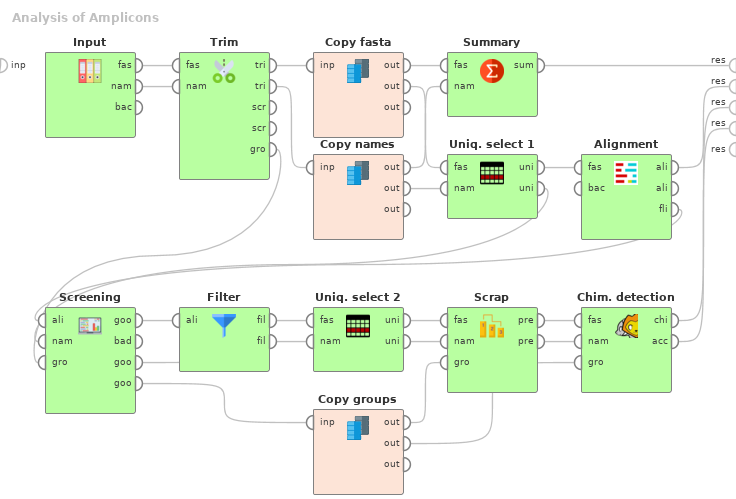
Процесс трансформации представляет собой анализ исходного кода Mothur, осуществляющийся при помощи алгоритмов, применяющих регулярные выражения для распознавания элементов структуры классов С++, задающих модули Mothur. Результаты анализа представляются в виде структур RDF (Resource description framework). Пример представления результата анализа модуля представлен на рисунке ...

|  |  |
| --- | --- |
| **@prefix** **dc:** <http://purl.org/dc/elements/1.1/> .  *# ......*  **@prefix** **xsd:** <http://www.w3.org/2001/XMLSchema#> .  **ngsp:spec** a **ngsp:Specification** ;  **ngsp:module** **mothur:chimera-ccode**,  *# ......*  **mothur:venn** .  *# ......*  **mothur:chimera-ccode** a **ngsp:Module** ;  **ngsp:outputPattern** [ a **cnt:Chars** ;  **ngsp:parameterName** *"type"* ;  *# ......*  [ **ngsp:patternString** *"[filename],[tag],..."* ;  **dc:identifier** *"accnos"* ],  [ **ngsp:patternString** *"[filename],mapinfo"* ;  **dc:identifier** *"mapinfo"* ] ;  **cnt:chars** *"getOutputPattern( // ... "* ;  **ngsp:parameter** **mothur:chimera-ccode-fasta-parameter**,  *# ......*  **mothur:chimera-ccode-window-parameter** ;  **dc:title** *"chimera-ccode"* ;  **dcterms:description** *"detect chimeric sequences"* ;  **schema:citation** *"..."* ; **schema:softwareHelp** *"..."* ;  **nco:websiteURL** <http://www.mothur.org/wiki/Chimera-ccode> ;  **v:category** *"Sequence Processing"* .  **mothur:chimera-ccode-fasta-parameter** a **ngsp:Parameter** ;  **ngsp:important** true ;  **ngsp:multipleSelectionAllowed** false ;  **ngsp:optionsDefault** *""* ;  **ngsp:outputTypes** [ a **oslc:AllowedValues** ;  **dc:identifier** *"accnos"*,*"chimera"*,*"mapinfo"* ] ;  **ngsp:outputTypesOrig** *"chimera-mapinfo-accnos"* ;  **ngsp:required** true ; **ngsp:type** **mothur:InputTypes** ;  **dc:title** *"fasta"* ; **schema:sku** 1 . |  |

* + 1. Результат анализа класса в исходном коде прикладного пакета Mothur: представление в виде RDF (слева), диаграмма SysML (справа)

Трансформация RDF-представления производится алгоритмами, реализованными в логической и объектно-ориентированной парадигмах программирования. Для обеспечения трансформации развит предложенный ранее подход применения логического программирования в качестве языка реализации процедур трансформации моделей (знаний о трансформации) в рамках «Модельно-управляемой архитектуры» (MDA, Model driven architecture). Использованная реализация трансформации в виде программы на языке LogTalk расширена новыми объектами, представляющими собой знания о программной целевой платформе, системы программирования Java.

Трансформация представляется как сценарий синтеза платформозависимой модели (PSM) программного обеспечения (ПО) (рисунок ЧЧЧ,а). Сценарий представляет собой объект – инкапсуляцию процедур трансформации, посылающий запросы объектам, представляющим платформонезависимую модель (PIM), и управляющий объектами, синтезирующими структуру PSM. PIM представлена в виде RDF-графа (рисунок ННН), а PIM – древовидная система блоков кода (объектов и литералов).

а) б)

* + 1. Общая структура процесса порождения исходного кода модуля (а) на основе данных анализа структуры прикладного пакета Mothur. Представление части вычислительного процесса анализа ампликонов (б) в виде визуальной диаграммы Rapidminer

Используемая архитектора инфраструктуры трансформации представлена на рисунке ЗЗЗ. Инфраструктура состоит из сервисов (большие блоки). Сервисы «CIM» (Computationally independent model) и «PIM» являются источниками информации, они заполняют базу данных «Сервера модели Mothur» исходными данными для трансформации. Сервер модели и «Сервер онтологий» предоставляют модулю трансформации («Инструмент MDA») исходные данные о структуре модулей Mothur, а также дополнительную информацию, например, из Wikipedia (DBPedia.org), позволяющую дополнить семантическими данными синтезируемую программу.

В рамках проекта используемая архитектура подверглась изменениям в части представления исходной информации CIM в виде результата анализа исходного кода Mothur. Также произведена настройка (в виде классов-адаптеров) блока Сервера моделей на разработанную концептуальную модель «ngs/ngsp» структуры Mothur. Данная модель позволяет задавать различные отношения, описывающий модуль Mothur (ngsp) и термины их предметной области NGS.



* + 1. Архитектура инфраструктуры анализа и порождения исходного кода  
       модулей Rapidminer

Одной из сопутствующих задач являлось исследование новых возможностей LogTalk для представления процедур трансформации. При реализации процедур трансформации использованы большинство предоставляемых LogTalk синтаксических структур и вариантов объектных иерархий. Особенно оказались полезными параметризованные объекты, категории и их иерархии прототипов. Параметризованные объекты позволяют задавать контекст выполнения метода без создания экземпляра класса. Второй необычной но полезной структурой являются категории – инкапсуляция реализаций логически связанных друг с другом методов – и композиционное программирование. Все статические объекты компилятор LogTalk (например, параметризованные объекты) преобразует непосредственно в Prolog, при этом накладные расходы на производительность кода не больше 1%.

На рисунке 999 представлен отрывок исходного кода объекта-адаптера запросов сценария синтеза исходного кода к исходным данным RDF-модели. Адаптер содержит один параметр – граф RDF и реализует три метода-запроса «class», «attribute» и «method», предназначенные для распознавания класса, списков его атрибутов и методов, соответственно.

В приведенном отрывке также представлен (в комментарии) пример запроса к данному адаптеру. Результатом запроса, как и в любой реализации Prolog, является множество ответов, представленных в виде наборов значений переменных. В данном примере в качестве ответа выступают наборы имя и идентификатор метода заданного идентификатором «ClassID» класса.

|  |
| --- |
| :- **object**(**query**(\_**RDF**)).  :- protected(xmi/1).  :- public([**class**/2, **attribute**/3, **method**/3]).  xmi(XMI) :- parameter(1, XMI).  **class**(Name, ID):- *% Распознавание класса в RDF*  ::xmi(XMI),  XMI::rdf(ID,rdf:type,uml:*'Class'*),  XMI::rdf(ID,rdfs:label, literal(Name)).  **attribute**(Name, ClassID, ID):- *% ...атрибута...*  ::xmi(XMI),  XMI::graph(G),  XMI::rdf(ClassID, G:ownedAttribute, ID),  *% XMI::rdf(ID, rdf:type, uml:'Property'),*  XMI::rdf(ID, rdfs:label, literal(Name)).  **method**(Name, ClassID, ID):- *% ...метода...*  ::xmi(XMI),  XMI::graph(G),  XMI::rdf(ClassID, G:ownedOperation, ID),  XMI::rdf(ID, rdfs:label, literal(Name)).  :- **end\_object.**  *%* ***query****(****mothur****)::****method****(Name,ClassID,MethodID).* |

* + 1. Параметризованный объект LogTalk, адаптирующий запросы к структурам   
       исходной RDF-модели

Процесс порождения исходного кода состоит из двух этапов. Первый – синтез модели генерируемого объекта, второй – собственно запись исходного кода объекта в определенное место в файле. На рисунке 555 представлен фрагмент генератора модуля Java. Известно, что согласно стандарту Java-1.8 в каждом модуле может находиться только один класс. В соответствии с данной спецификацией генератор модуля состоит из двух методов: генератора преамбулы и генератора класса. Генератор класса реализован в отдельном модуле. Кроме методов генерации модели исходного кода генератор модуля содержит вспомогательные методы, используемые на этапе порождения исходного кода конфигурационных файлов Rapidminer. Например, метод «module\_icon\_name» синтезирует имя файла, в котором храниться пиктограмма модуля, на основе имени класса.

|  |
| --- |
| :- **object**(mothur\_module,  **specializes**(java\_module)).  :- public(**preamble**/0). *% генерирует начало модуля.*  **preamble**:-  ^^preamble,  ::set\_package(*'com.rapidminer.ngs.operator'*),  ::add\_skip(1),  **create\_object**(Imports, [instantiates(java\_import)],[],[]),  ::set\_block(imports(Imports)),  ::append(Imports),  ::add\_skip(1),  **create\_object**(ClassDef, [instantiates(mothur\_class)],[],[]),  ClassDef::preamble,  ::set\_block(class(ClassDef)),  ::append(ClassDef),  Imports::add(*'java.util.List'*),  Imports::add(*'com.rapidminer.operator.OperatorDescription'*),  Imports::add(*'com.rapidminer.operator.ports.InputPort'*),  *% . . . . . . .*  *% Указание родительского класса генерируемому классу.*  ClassDef::extends(*'MothurGeneratedOperator'*).  :- public(**module\_class**/2).  **module\_class**(Class, Name):- *% Генератор класса модуля*  ::current\_block(class(Class)),  Class::class\_name(Name).  :- public(**module\_name**/1).  **module\_name**(String):-  ::module\_class(\_,Name),  writef::swritef(String,*'%w.java'*, [Name]).  *% . . . . . . . . .*  :- public(**module\_icon\_name**/1).  **module\_icon\_name**(String):-  ::module\_class(Class,\_),  Class::item(name(Name)),  downcase\_atom(Name, DName),  writef::swritef(String, *'icon-operator-%w.png'*, [DName]).  :- **end\_object.** |

* + 1. Генератор модуля Rapidminer

Процесс записи синтезируемого кода реализуется предикатами метода «renderitem/2», который дополняется новыми знаниями при наследовании классов-генераторов. На рисунке 000 представлен пример такого дополнения. Метод «renderitem» дополнен правилом синтеза исходного кода элемента «imports». Основная процедура порождения исходного кода, как правило, последовательно обходит в глубину структуру модели кода «code\_block» и запускает методы «renderitem» для каждого найденного элемента.

|  |
| --- |
| :- **object**(java\_module,  **specializes**(code\_block)).  *% . . . . . . .*  **renderitem**(package(Name), String):-!,  root::iswritef(String, *'package %w;'*, [Name]).  **renderitem**(imports(Object), String):-!,  Object::render(String).  **renderitem**(Item, String):-  ^^renderitem(Item, String). % Возов в суперклассе  *% . . . . .*  :- public(**current\_imports**/1).  current\_imports(Object):-  ::item(imports(Object)).  :- public(**preamble**/0).  preamble. *% Do nothing*  :- **end\_object.** |

* + 1. Реализация экспорта исходного кода

Система Rapidmier Studio поддерживает блоки обработки данных в виде расширений системы (plug-ins). Каждый модуль расширения помимо генерируемых классов содержит метаописания, позволяющие классы интегрировать в визуальную среду, а также снабдить все информационные единицы текстами контекстно помощи на разных языках. Метаописания также генерируются автоматически, исходя из структуры класса и исходных RDF-данных. В результате окончательный вариант визуального представления процедур Mothur представляет собой набор блоков из которых строиться модель вычислительного процесса, ример которой представлен на рисунке ИИИ,б.

Таким образом, в результате разработки данного направления создана и формализована система знаний, при помощи которой по результатам анализа исходного кода Mothur осуществляется порождение исходного кода Java-модулей для системы визуального проектирования вычислительных процессов Rapidminer. В виде блоков Rapidminer представлены 143 процедуры обработки данных NGS. На следующем этапе исследований по данному направлению необходимо разработать генераторы визуального представления модулей вычислительных сред NGS, таких как QIIME, QIIME2, Usearch и др., что позволит решить две задачи. Первая - это расширить методический комплекс алгоритмами других прикладных пакетов, вторая - это актуализировать процесс проектирования вычислений на текущее состояния отрасли NGS.

**Заключение**

Таким образом, для поддержки исследований в метагеномном анализе, который позволяет описывать микробные сообщества с ранее недоступной точностью, на данном этапе разработана система формализованных знаний, позволяющих преобразовывать описания струкр модулей анализа NGS в визуальные модели. На основе этих моделей на следующих этапах проекта будут реализованы программы визуального программирования процедур анализа ампликонов, модули обеспечения хранения исходных и промежуточных данных в облачном хранилище, интегрировать процедуры и данных, организованные разными исследованиями, публиковать семантически-размеченные данные в Интернете, а также разрабатываться методики автоматизации проектирования распределенных вычислительных процессов на вычислительных кластерах.

**Список использованных источников**

1. Paulo Moura. A Portable and Efficient Implementation of Coinductive Logic Programming // Proceedings of the Fifteenth International Symposium on Practical Aspects of Declarative Languages (PADL 2013), 2013. Springer LNCS 7752.

**Публикации**

1. Evgeny Cherkashin, Alexey Kopaygorodsky, Ljubica Kazi, Alexey Shigarov, and Viacheslav Paramonov, Model driven architecture implementation using linked data// Procs. of 24th International Conference, ICIST 2018, Vilnius, Lithuania, October 4-6, 2018. P. 412-423.
   1. ЗАКЛЮЧЕНИЕ

Первый этап проекта является начальным этапом проведения научных исследований и направлен, в основном, на сбор методического материала, построения информационных и математических моделей общего вида, а также сбор исходной информации для решения задач проекта в рамках основной идеи – реализации системного подхода к исследованию озера Байкал, основывающегося на анализе и интерпретации данных секвенирования нового поколения (NGS). Проведение научных исследований в рамках предложенного подхода, с одной стороны, позволит интегрировать имеющиеся данные и получаемые результаты в мировую информационную среду и среду знаний о морских ресурсах планеты Земля, а с другой – разработать автоматизированную процедуру анализа состояния экосистем природных объектов, начиная с этапа планирования эксперимента и сбора проб до выявления закономерностей между продуцентами, синтеза управляющих воздействий, вводящих природный объект в требуемое состояние.

В 2017 году получены следующие результаты:

1. По предметному направлению NGS – осуществлен выбор метода экстракции ДНК, который протестирован на имеющихся данных и новых пробах, метод показал свою применимость при использовании существующего оборудования ЛИН СО РАН и его партнеров.
2. Разработана стратегия фильтрации данных секвенирования, реализующая так называемый консервативный подход к удалению некорректных исходных данных; на тестовых данных показано, что использование данной стратегии предотвращает эффект *обеднения таксономического разнообразия*: предложенный подход позволяет получать более достоверные результаты NGS.
3. Впервые проведен метагеномный анализ геномов микроорганизмов из трофогенного слоя (5 и 20 м) Южного Байкала в подледный период, показано, что большинство восстановленных геномов являются эндемиками озера Байкал.
4. Изолированы два термофильных, факультативно анаэробных штамма микроорганизмов, его физиолого-биохимические свойства не отличаются от свойств типовых термофильных штаммов этого рода, но имеющем обособленное филогенетическое положение внутри кластера последовательностей типовых видов (рисунок 4.2), что дает основание отнести данный штамм к новому виду микроорганизмов.
5. Осуществлен анализ сложившейся практики хранения и обработки данных в ЛИН СО РАН и существующего программного обеспечения автоматизации разных этапов NGS, в результате чего принято решение о реализации облачного хранилища данных, позволяющего автоматизировать различные этапы NGS; разработанные функциональные блоки хранилища протестированы на имеющихся данных, показано соответствие разработанных структур задачам предметных исследований.
6. Разработан подход к визуализации процесса секвенирования на основе программирования потоков данных, показано, что операции прикладного пакета Mothur представимы элементами процесса обработки потоков данных, разработан начальный набор операций для визуального программирования процессов в NGS.
7. Произведен анализ современных средств интеграции данных исследований, показано, что использование технологий открытых связанных данных позволяет решать основные задачи интеграции, разработано ядро сервиса предоставления (публикации) данных и результатов исследований.
8. Создана процедура идентификации динамической модели микробиома озера Байкал, интерпретирующая корреляционные связи между представимостью (численностью) микроорганизмов, их химических взаимодействий и параметрами окружающей среды; создан прототип программы поддержки процесса моделирования.

Результаты, полученные в проекте в 2017 представляют собой реализацию современной метеорологии исследований водоемов, широко используемых в мировой практике проведения научных исследований, но впервые примененных к новому объекту исследования – оз. Байкал. Полученные результаты носят мультидисциплинарный характер и на этапе 2017 года относятся к стадии сбора и первичной обработки информации, а также к совершенствованию существующих процессов секвенирования нового поколения. Создаваемое методическое, программное и инфраструктурное обеспечение NGS позволит успешно пройти следующие стадии проекта.

* 1. СПИСОК ИПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Элементы списка литературы оформляются стилем **Bibitem**, он соответствует ГОСТУ, и мне тоже не нравится.

Примеры оформления из ГОСТ

1. Гуреев В.Н.. Маэов НА Использование библиометрии для оценки значимости журналов в научных бибшотеках (обзор)// Научно-техническая информация. Сер. 1. —2015. — N9 2. —С. 8—19.
2. Колкова Н.И.. Скипор И.Л. Термин осистеыа предметной области «электронные информационные ресурсы»: взгляд с позиций теории и практики // Научн. и техн. б-ки. — 2016. — N9 7. — С. 24—41.

Книги, монографии:

1. Земсхов А.И., Шрайберг Я.Л. Электронные библиотеки: учебник для вузов. — М.: Пиберея. 2003. — 351 с.
2. Костюк К.Н. Книга в новой медицинской среде. — М.: Директ-Медиа. 2015. —430 с.

Тезисы докладов, материалы конференций:

1. Пеготин Е.Ю. Организация метаданных в хранилище данных // Научный поиск. Технические наухи: Мате-риалы 3-й нзуч. конф. аспирантов и докторантов / огв. за вып. СД. Ваулин: Юж.-Урал. гос. ун-т. Т. 2. — Челябинск: Издательский центр ЮУрГУ. 2011. — С. 128—132.
2. Ангопольский А.Б. Система метаданных в электронных библиотеках II Библиотеки и ассоциии в меняющемся мире: Новые технологии и новые формы сотрудничества: Тр. 8-й Междунар. конф. «Крым-2001» / г. Судак,(июнь 2001 г.). — Т. 1. — М.. 2001. — С. 287—298.
3. Парфенова С.Л., Гришакина Е.Г., Золотарев Д.В. 4-я Международная научно-практическая конференция «Научное издание международного уровня — 2015: современные тенденции в мировой практике редактирования, издания и оценхи научных публикаций» // Наука. Инновации. Образование. — 2015. — N9 17. — С. 241—252.

Патентная документация согласно стандарту ВОИС:

1. BY (код страны) 18875 (N9 патентного документа) С1 (код вида документа). 2010 (дата публикации).

Электронные ресурсы:

1. Статистические показатели российского книгоиздания в 2006 г.: цифры и рейтинги (Электронный ресурс]. —2006. — URL: ht1pyrbookhambw.ru/stat\_2006.htm (дата обращения 12.03.2009).
2. Прогноз научно-технологического развития Российской Федерации на период до 2030 года. — URL: http://govemment.ru/media/fiies/41d4b737638891da21&4/pdf (дата обращения 15.11.2016).
3. Web of Science. — URL: <http://apps.webofknowledge.com/> (дата обращения 15.11.2016).

Нормативные документы:

1. ГОСТ 7.0.96—2016 Система стандартов по информации, библиотечному и издательскому делу. Электронные библиотеки. Основные виды. Структура. Технология формирования. — М.: Сгакдартинформ. 2016. — 16 с.
2. Приказ Минобразования РФ от 19 декабря 2013 г. N9 1367 «Об утверждении Порядка организации и осуществления образовательной деятельности по образовательным программам высшего образования — про- граммам бакалавриата, программам специалитега. программам магистратуры». — URL: <http://www>.consullant.ru/document/cons\_doc\_LAW\_159671/(дата обращения: 04.08.2016).
3. ISO 25964-1:2011. Information and documentation —Thesauri and interoperabdity with other vocabularies — Part 1: Thesauri for information retrieval. — URL: <http://www.iso.org/iso/catalogue> detail.htm?csnumber=53657 (дата обраще- ния: 20.10.2016).
4. Karsenti et al. A Holistic Approach to Marine Eco-Systems Biology [Текст] // PLoS Biol. 2011. Vol. 9. № 10. P. e1001177;
5. Karsenti Towards an ‘Oceans Systems Biology’ [Текст] // Mol. Syst. Biol. 2012. Vol. 8. P. 575
6. Gerasimidis K., Bertz M., Quince C., Brunner K., Bruce A., Combet E., Calus S., Loman N., Jiaz U.Z. The effect of DNA extraction methodology on gut microbiota research applications [Текст] // [BMC Res Notes](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4960752/). 2016; 9: 365.doi: [10.1186/s13104-016-2171-7](https://dx.doi.org/10.1186/s13104-016-2171-7).
7. Wen Y., Xiao F., Wang C.,Wang Z. The impact of different methods of DNA extraction on microbial community measures of BALF samples based on metagenomic data [Текст] // [Am J Transl Res](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4858570/). 2016; 8(3): 1412–1425.
8. Fouhy F., Clooney A.G., Starton C., Claesson M., Cotter P.D. 16S rRNA gene sequencing of mock microbial populations- impact of DNA extraction method, primer choice and sequencing platform [Текст] // [BMC Microbiol](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4921037/). 2016; 16: 123.doi: [10.1186/s12866-016-0738-z](https://dx.doi.org/10.1186/s12866-016-0738-z).
9. Rusch D.B., Halpern A.L., Sutton G., Haidelberg K.B., Williamson S., Yooseph S., Yooseph S., Wu D., Eisen J.A., Hoffman J.M., Remington K., Beeson K., Tran B., Smith H. The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific [Текст] // Plos Biology, 2007, <https://doi.org/10.1371/journal.pbio.0050077>
10. Allen L.Z., Allen E.E., Badger J.B., McCrow J.P., Paulsen I.T., Elbourne L.DH., Thiagarajan M., Rusch D.B., Nealson ЛюРюб Williamson S., Venter J.C., Allen A.E. Influence of nutrients and currents on the genomic composition of microbes across an upwelling mosaic [Текст]// [ISME J](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3379637/). 2012; 6(7): 1403–1414. doi: [10.1038/ismej.2011.201](https://dx.doi.org/10.1038/ismej.2011.201).
11. Smith M.W., Allen L.Z., Allen A. E., Simon H.M., Contrasting genomic properties of free-living and particle-attached microbial assemblages within a coastal ecosystem [Текст] // Front. Microbiol., <https://doi.org/10.3389/fmicb.2013.00120>.
12. Шубенкова О.В., Земская Т.И., Черницына С.М., Хлыстов О.М., Трибой Т.И. Первые результаты исследования филогенетического разнообразия микроорганизмов осадков Южного Байкала в районе приповерхностного залегания гидратов метана [Текст] // Микробиология. 2005. Т. 74. № 3.С. 370–377.
13. Petrosino, J. F., Highlander, S., Luna, R. A., Gibbs, R. A., Versalovic, J.*.* Metagenomic pyrosequencing and microbial identification [Текст] //Clinical chemistry. – 2009. – Т. 55. – №. 5. – С. 856-866.
14. Kim, M., Lee, K. H., Yoon, S. W., Kim, B. S., Chun, J., Yi, H. (2013). Analytical tools and databases for metagenomics in the next-generation sequencing era. [Текст] // *Genomics & informatics*, *11*(3), 102-113.
15. Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., ... & Gu, Y (2012)*.* A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers [Текст] //BMC genomics. – 2012. – Т. 13. – №. 1. – С. 341.
16. Magoč, T., & Salzberg, S. L. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. [Текст] //*Bioinformatics*, *27*(21), 2957-2963.
17. Fosso, B., Santamaria, M., Marzano, M., Alonso-Alemany, D., Valiente, G., Donvito, G., ... & Pesole, G. (2015). BioMaS: a modular pipeline for Bioinformatic analysis of Metagenomic AmpliconS. [Текст] //*BMC bioinformatics*, *16*(1), 203.
18. Tennant, R. K., Sambles, C. M., Diffey, G. E., , K. A., & Love, J. (2017). Metagenomic Analysis of Silage. [Текст] // *Journal of visualized experiments: JoVE*, (119)
19. Katoh, K., & Toh, H. (2010). Parallelization of the MAFFT multiple sequence alignment program. [Текст] //*Bioinformatics*, *26*(15), 1899-1900.
20. Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Sahl, J. W. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. [Текст] //*Applied and environmental microbiology*, *75*(23), 7537-7541.
21. Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Glöckner, F. O. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. [Текст] //*Nucleic acids research*, *41*(D1), D590-D596.
22. Chun, J., Kim, K. Y., Lee, J. H., & Choi, Y. (2010). The analysis of oral microbial communities of wild-type and toll-like receptor 2-deficient mice using a 454 GS FLX Titanium pyrosequencer. [Текст] //*BMC microbiology*, *10*(1), 101.
23. Smith, E. P., & van Belle, G. (1984). Nonparametric estimation of species richness. [Текст] //*Biometrics*, 119-129.
24. Dixon, P. (2003). VEGAN, a package of R functions for community ecology. [Текст] //*Journal of Vegetation Science*, *14*(6), 927-930
25. Giovannoni SJ, Thrash JC, Temperton B. 2014. Implications of streamlining theory for microbial ecology. [Текст] //The ISME journal 8:1553.
26. Morris RM, Rappé MS, Connon SA, Vergin KL. 2002. SAR11 clade dominates ocean surface bacterioplankton communities. [Текст] //Nature 420:806
27. Daims H, Lebedeva EV, Pjevac P, Han P, Herbold C, Albertsen M, Jehmlich N, Palatinszky M, Vierheilig J, Bulaev A. 2015. Complete nitrification by Nitrospira bacteria. [Текст] //Nature 528:504.
28. Weiss R, Carmack E, Koropalov V. 1991. Deep-water renewal and biological production. [Текст] //Nature 349:665.
29. Zeng Y, Feng F, Medová H, Dean J, Koblížek M. 2014. Functional type 2 photosynthetic reaction centers found in the rare bacterial phylum Gemmatimonadetes. [Текст] //Proceedings 884 of the of Sciences 111:7795-7800.
30. Hugerth LW, Larsson J, Alneberg J, Lindh MV, Legrand C, Pinhassi J, Andersson AF. 2015. Metagenome-assembled genomes uncover a global brackish microbiome. [Текст] //Genome biology 16:1-18.
31. Mehrshad M, Amoozegar MA, Ghai R, Fazeli SAS, Rodriguez-Valera F. 2016. Genome reconstruction from metagenomic datasets reveals novel microbes in the brackish waters of the Caspian Sea. [Текст] //Applied and environmental microbiology:AEM. 03381-15.
32. Brock, T.D. (1978). Thermophilic microoganisms and life at high temperatures. US: Springer-Verlag. Bonch-Osmolovskaya, E.A. (2005). Phylogenetic and metabolic diversity of thermophilic prokaryotes with different types of anaerobic respiration. Microbial diversity: current perspectives and potential application: International Publishing House [Текст].
33. Rozanov, A.S., Ivanisenko, T.V., Bryanskaya, A.V., Shekhovtsov, S.V., Logacheva, M.D., Saik, O.V., Malup, T.K., Demenkov, P.S., Goryachkovskaya, T.N., Ivanisenko, V.A., Peltek, S.E. (2014). Bioinformatic analysis of the genome of the Geobacillus stearothermophilus 22 strain isolated from the Garga, Baikal region. [Текст] //Russian J. Genetics: Applied Research, 4, 267–272.
34. Hollander M., Wolfe D.A. (1973), Nonparametric Statistical Methods [Текст]: John Wiley & Sons. Pages 185–194 (Kendall and Spearman tests).
35. Patrick Royston (1995) Remark AS R94: A remark on Algorithm AS 181: The W test for normality. [Текст] //Applied Statistics, 44, 547–551.
36. Gregory R. Warnes, Ben Bolker, Lodewijk Bonebakker, Robert Gentleman, Wolfgang Huber Andy Liaw, Thomas Lumley, Martin Maechler, Arni Magnusson, Steffen Moeller, Marc Schwartz and Bill Venables (2015). gplots: Various R Programming Tools for Plotting Data. R package version 2.17.0. [Электронный ресурс] <http://CRAN.R-project.org/package=gplots>
37. Tesson B, Hildebrand M (2010) Extensive and intimate association of the cytoskeleton with forming silica in diatoms: control over patterning on the meso- and micro-scale. [Электронный ресурс] PLoS ONE 5:e14300.
    1. Приложение А  
        Публикации по теме проекта за 2017 год
38. Башенхаева М.В., Захарова Ю.П., Галачьянц Ю.П., Ханаев И.В., Лихошвай Е.В. Сравнительный анализ бактериальных сообществ озера Байкал в подледный период и период открытой воды [Текст] // Acta Naturae. – 2017. – Т. 9, № 1. – С. 30. WoS+**РИНЦ + 4.1.2.**
39. Букин Ю.С., Галачьянц Ю.П. Процедура тримминга метагеномных данных, полученных путем высокопроизводительного секвенирования ампликонов [Текст] // Acta Naturae. – 2017. – Т. 9, № 1. – С. 19. WoS+**РИНЦ + 4.1.2.**
40. Земская Т.И., Ломакина А.В., Захаренко А.С., Хальзов И.А., Черницина С.М., Шубенкова О.В., Павлова О.Н., Букин С.В., Галачьянц Ю.П., Морозов И.В. Метагеномный анализ микроюных сообществ для исследования цикла метана в озере Байкал [Текст] // Acta Naturae. – 2017. – Т. 9, № 1. – С. 33. WoS+ **РИНЦ + 4.1.2. и 0007**
41. Лихошвай Е.В. Перспективы применения NGS для структурно-функциональной характеристики водных экосистем [Текст] // Acta Naturae. – 2017. – Т. 9, № 1. – С. 36. WoS+**РИНЦ + 4.1.2.**
42. Малков Ф.С., Галачьянц Ю.П., Шигаров А.О., Морозов А.А., Михайлов И.С., Ломакина А.В., Захаренко А.С. Управление данными в исследовании микробиома оз. Байкал [Текст] // Тезисы докладов 18-й Всеросс. конф. молодых ученых по математическому моделированию и информационным технологиям. Иркутск, 2017. c. 82. Работа выполнена в рамках интеграционного проекта Иркутского научного центра СО РАН № 4.1.2.
43. Малков Ф.С., Черкашин Е.А., Шигаров А.О. Перспективы применения системы управления рабочим процессом Taverna в задачах обработки метагеномных данных при исследовании микробиома озера Байкал [Текст] // Материалы конференции «Ляпуновские чтения». 2017. С. 32.
44. Морозов А.А., Шигаров А.О., Малков Ф.С., Паскал К.К., Черкашин Е.А., Лихошвай Е.В. Информационно-вычислительная инфраструктура для поддержки метагеномного анализа [Текст] // Вестник ИГУ. – 2017. **РИНЦ +** **ИНЦ + 4.1.2.**
45. Павлова О.Н., Букин С.В., Горшков А.Г., Ханаева Т.А., Земская Т.И. Микроорганизмы озера Байкал: от психрофильных углеводородокисляющих аэробов до термофильных миксотрофов [Текст] // 1-й Российский микробиологический конгресс. 17–18 октября 2017 г. Пущино, Россия: ООО Пущино: «ИД «Вода: химия и экология». – 2017. – С. 68–69. **4.1.2.**
46. Ханаева Т.А., Павлова О.Н., Черницына С.М., Хальзов И.А., Хабуев А.В., Никонова А.А., Новикова А.С., Земская Т.И. Термофильная факультативно анаэробная бактерия р. *Geobacillus* из донных осадков озера Байкал. [Текст] //Acta Biologica Sibirica, 2017/ **РИНЦ +ИНЦ + 4.1.2.**
47. Черкашин Е.А., Шигаров А.О., Орлова И.В., Михайлов И.С. Использование технологий Linked Open Data при публикации текстовых документов [Текст] // Материалы Всеросс. конф. «Знания-Онтологии-Теории». Новосибирск, 2017. Т. 2, с. 138-147. **РИНЦ+** **ИНЦ + 4.1.2.**
48. Bukin Yu. S., Buzoleva L.S., Golozubova Y.S., Galachyants Yu. P. New procedure of raw Illumina MiSeq data filtering for the amplicon metagenomic libraries [Текст] // Mathematical Biology and Bioinformatics. – under rev. Scopus + **ИНЦ** +**4.1.2.**
49. Cabello-Yeves P., Zemskaya T., Rosselli R., Coutinho F., Zakharenko A., Blinov V., Rodriguez-Valera F. Genomes of novel microbial lineages assembled from the sub-ice waters of Lake Baikal. [Электронный ресурс]//Accepted manuscript posted online 27 October 2017. 2018. #84:1 doi: 10.1128/AEM.02132-17 WoS **4.1.2.**
50. Cherkashin E., Shigarov A., Orlova I., Mikhailov I. Authoring and Publishing Text Documents by means of Linked Open Data Technologies [Электронный ресурс]// In Procs. Int. Conf. on Applied Internet and Information Technologies., 2017. pp. 98-109. URL: <http://tfzr.rs/aiit/files/ProceedingsAIIT2017.pdf> **+** **ИНЦ + 4.1.2.**
51. Cherkashin E., Shigarov A., Malkov F., Pascal K., Morozov A. An Environment for Metagenomic Analysis [Электронный ресурс]// In Proc. Int. Conf. on Applied Internet and Information Technologies. , , 2017. pp. 110-117. URL: [http://tfzr.rs/aiit/files/](http://tfzr.rs/aiit/files/ProceedingsAIIT2017.pdf)**+** **ИНЦ + 4.1.2.**
52. Lomakina A.V., Chernitsyna S.M., Pogodaeva T.V., Zakharenko A.S., Zmskaya T.I. Diversity of microbial communities in zones with a contrast mineralization of pore waters on Lake Baikal [Электронный ресурс]// 13th international conference on salt lake research (ICSLR 2017). August 21–25, , – 2017. – P. 92–93. **РИНЦ +ИНЦ + 4.1.2.**
    1. Приложение Б  
       Информация о количестве статей в виде таблицы

|  |  |  |
| --- | --- | --- |
| Индикатор | Ед. измерения | 2018 |
| Количество публикаций в ведущих российских и международных журналах по результатам исследований, полученных в процессе реализации проекта | единиц | 10 |
| Количество публикаций в мировых научных журналах, индексируемых в базе данных «Сеть науки» (WEB of Science) и/или Scopus | единиц | 2 |
| РИНЦ (исключая WEB of Science и Scopus) | единиц | 8 |

* 1. ПРИЛОЖЕНИЕ В   
     Этапы выполнения проекта

*Этап 1* (2018 г.)

1. Отработка методов выделения суммарной ДНК из разных источников (вода, донные отложения) для секвенирования библиотек ампликонов фрагментов маркерных генов 16S (прокариоты) и 18S (эукариоты) рРНК и фрагментных метагеномных библиотек с целью получения информации о таксономической структуре эукариотических организмов в сообществах, соответствующей данным микроскопических методов анализа.
2. Отбор проб из фотического слоя и донных осадков (до 100), выделение ДНК, секвенирование библиотек ампликонов 16S и 18S рРНК (общий объем секвенирования – 1-2 запуска Illumina MiSeq или аналогичный), таксономический анализ сообществ.
3. Подготовка материалов для регистрации выделенных в культуру бактерий во Всероссийскую коллекцию микроорганизмов (ВКМ) (г. Пущино) и немецкую коллекцию микроорганизмов и клеточных культур, г. Лейбниц, Германия (DSMZ).
4. Отработка алгоритмов биоинформатического анализа результатов секвенирования фрагментных метагеномных библиотек.
5. Разработка онтологии и технологии унифицированного доступа к данным результатов исследований на основе Linked-Data и современных стандартов представления данных о биомах. Интеграция хранилища данных проекта в мировую информационную среду.
6. Анализ современного программного обеспечения обработки BD, используемого в приложениях NGS, в том числе, удалении адаптеров и химер, проверке качества данных, таксономической идентификации, картировании прочтений, сборки *de novo*, анализе ампликонов, реконструкции метаболических путей и аннотации последовательностей. Выбор оптимального ПО для обработки данных
7. Подготовка технических предложений по проектированию и реализации ИОБД, предназначенной для архивирования, управления, курирования, анализа и визуализации данных, получаемых в результате исследования микробиома биоинформационными технологиями и NGS.
8. Проектирование схемы Хранилища Больших Данных для обеспечения управления данными, используемыми при исследовании микробиома Байкала, а также веб-публикации для поиска нуклеотидных и аминокислотных последовательностей по заданным таксонам и параметрам среды обитания.
9. Для построения полисистемы динамических моделей выделение сквозных концептов, имеющих качественный характер, и базовых взаимосвязей между ними. Проведение серии пробных идентификаций параметров.