**Среда поддержки научных исследований в секвенировании нового поколения (NGS)**

Исследования по данному направлению направлены на решение проблемы проектирования и реализации распределенной программной среды для создания организационных, информационных и вычислительных ресурсов проведения научных микробиологических исследований на основе метагеномного анализа. Построена обобщенная модель предметной области системного уровня и выделены требования к разрабатываемой среде, задачи, требующие решения; построена и реализована модель вычислительного процесса анализа ампликонов в виде графа управления вычислительным процессом в среде Rapid Miner. Разрабатываемая программная среда позволит решать задачу построения информационного портала для обработки метагеномных данных и представления результатов научному сообществу.

Для обеспечения  метагеномных исследований требуются значительные вычислительные ресурсы, а также участие квалифицированного биоинформатика в обработке и интерпретации данных. 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](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.

Существуют два основных вида метагеномных исследований. Первый, более простой, называется анализом ампликонов. При этом амплифицируется и секвенируется определённый таксономический маркер, универсальный для исследуемых видов. Прочтения, полученные при секвенировании выделенной из исследуемого образца ДНК, сравниваются с существующими 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. На основе метагеномных данных устанавливаются метаболические взаимодействия в отдельных микробиомах, опираясь на базы данных ePGDBs (environmental pathway/genome databases). В некоторых работах из метагеномных наборов прочтений удавалось выделить полные геномы отдельных видов.

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. 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. Вычислительный процесс представляет собой, как правило, последовательность запусков вычислительных утилит с передачей им параметров в виде набора файлов. Вычислительный процесс представляет собой сетевой граф элементарных операций, который строится на основе передаваемой между модулями информации. Циклические структуры редки, причем они, в основном, порождаются в результате анализа промежуточных данных и результатов специалистами-предметниками: при неудовлетворительном качестве тот или иной модуль перезапускается человеком вручную с новыми параметрами.

Обработка 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. Поэтому предложенная и реализуемая концепция организации 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. В результате в среде строятся специальные варианты вычислительных процессов, решающих задачи, поставленные биологом. Необходимо также обеспечить также организованное облачного хранения промежуточных данных и результатов. На этапе 2017 года в проекте решалась задача исследования предметной области вычислительного процесса и ее декомпозиция на классы задач (функциональное моделирование). В результате такого моделирования выделены классы задач, которые, затем, представлены в виде программных модулей. Из модулей формируются сценарии решения задач, сетевые графы модулей, связанные друг с другом передачей данных.

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. 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 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 Рис. 1 представлен пример (сценарий) использования разрабатываемой в проекте системы управления вычислительным процессом анализа ампликонов.

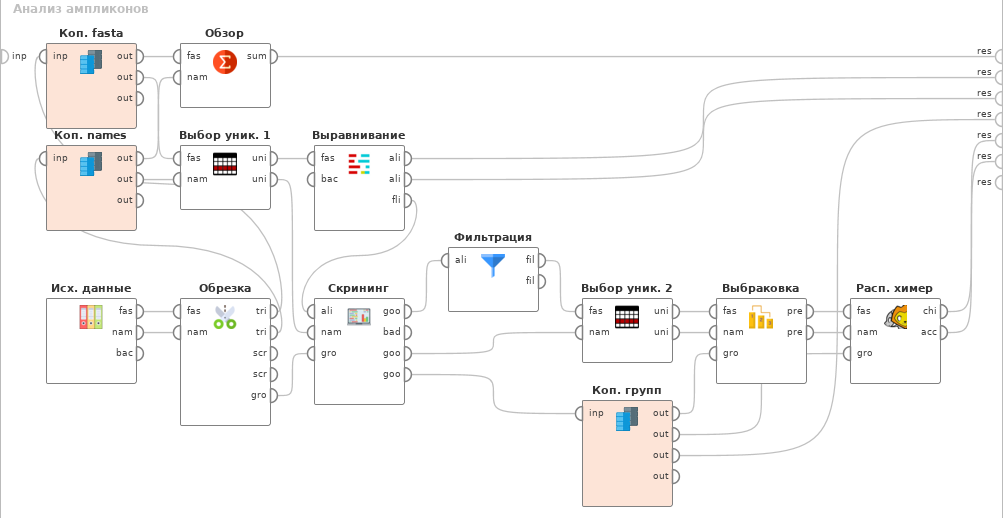


Рис. 1. Сетевой граф начальной стадии процесса анализа ампликонов

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

1. формирование исследовательского проекта в виде папки (директории), содержащей данные секвенирования (модуль “Исх. данные”);
2. обрезка данных (прочтений) (модуль “Обрезка”);
3. модуль “Обзор” используется для визуального анализа качества результатов предыдущих шагов;
4. сокращение объема входных данных за счет удаления незначащей информации, например, дублирующих последовательностей (модули “Выбор уник.”);
5. выравнивание последовательностей на референсные базы данных (модуль “Выравнивание”);
6. фильтрация последовательностей по заданным критериям (модуль “Скрининг”);
7. удаление колонок выравнивания по заданным критериям, например, пустых колонок (модуль “Фильтрация”);
8. удаление последовательностей, содержащих ошибки секвенирования (модуль “Выбраковка”);
9. 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, индивидуального для каждого модуля. 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.

Модули RapidMiner, отождествляемые с операциями NGS, разрабатываются на языке программирования Java. Выбор этого языка обусловлен тем, что сам редактор RapidMiner Studio реализован а системе программирования Java, а стандартизированных интерфейсов к другим системам программирования разработчиками еще не было сделано. В связи с этим процесс программирования операции представляет собой

1. Создание унаследованного класса, реализующего операцию: задание портов ввода и вывода, задание перечня параметров модуля, преобразование данных портов в вызов утилиты Motur, декодирование результатов работы программы, анализ ошибок времени исполнения Motur.
2. Описание интерфейса класса в виде XML-файла, читаемого редактором RapidMiner Studio.
3. Задание положения модуля в структуре модулей редактора.
4. Разработку пиктограммы операции, отображаемой внутри блока и символизирующей сущность операции.

Дальнейшим направлением развития специализированной версии RapidMiner является а) разработка системы интерпретации метаописаний операций программы Mothur; б) реализация новых операций; в) интеграция с облачным хранилищем (базой данных), разрабатываемым в рамках проекта; г) создание интерфейса пользователя для специалистов-биологов, позволяющей им самостоятельно проводить исследования и в некоторой степени влиять на вычислительный процесс; д) разработка специализированных приложений и библиотек операций, решающих специальные задачи.

Таким образом, для поддержки исследований в метагеномном анализе, который позволяет описывать микробные сообщества с ранее недоступной точностью, положено начало разработки средств автоматизации организации исследований, направленных на обеспечение доступа к значительным вычислительные мощностям и мощным алгоритмам анализа ампликонов специалисту-биологу без значительного участия биоинформатиков. Данное свойство системы обеспечивается при помощи визуального представления операций, обобщением функций программного обеспечения и разработке интерфейсов пользователя, ориентированных на специалистов-предметников. В целом, разрабатываемая среда поддержки научных исследования позволит предоставить исследователям удобный интерфейс анализа и обеспечить хранение в унифицированной форме исходных данных и результатов анализа.

**Публикации по теме**

1. (ИГУ в печати)
2. E.Cherkashin, A.Shigarov, F.Malkov, K.Pascal, A.Morozov. An Environment for Metagenomic Analysis. Procs. of International Conference on Applied Internet and Information Technologies (ICAIIT) 2017. Zrenjanin, October 5-6, Serbia, 2017. p. 110-117. URL: [http://tfzr.rs/aiit/files/](http://tfzr.rs/aiit/files/ProceedingsAIIT2017.pdf) (access date: 18.10.2017) ISBN 978-86-7672-304-1