

Санкт-Петербургский национальный исследовательский университет
информационных технологий, механики и оптики
Факультет информационных технологий и программирования
Кафедра компьютерных технологий

**Реализация эффективного взаимодействия между
платформой для анализа экспрессии генов Morpheus и
библиотекой вычислительных методов R/Bioconductor**

Зенкова Д.М.

Научный руководитель: Сергушичев А. А.

Санкт-Петербург
2017

ОГЛАВЛЕНИЕ

Стр.

ВВЕДЕНИЕ	6
ГЛАВА 1. ОБЗОР ПРЕДМЕТНОЙ ОБЛАСТИ	7
1.1. Биоинформатика	7
1.1.1. Анализ экспрессии генов.....	7
1.1.2. Используемые методы	7
1.2. Существующие решения для анализа экспрессии генов ...	8
1.2.1. GENE-E	8
1.2.2. morpheus.js	8
1.2.3. R/Bioconductor.....	9
1.3. Инструменты, которые могут быть применены	9
1.3.1. Язык R и библиотека Bioconductor.....	9
1.3.2. JavaScript	9
1.3.3. R shiny	9
1.3.4. OpenCPU	9
1.3.5. Gene Expression Omnibus.....	10
1.3.6. Docker	10
1.3.7. JSON.....	10
1.3.8. Protocol Buffers	10
1.3.9. Apache2.....	10
1.3.10. HTML.....	10
1.4. Постановка задачи	10
1.4.1. Цель работы	10
1.4.2. Основные задачи	11
1.4.3. Требования к веб-приложению phantasus	11
Выводы по главе 1.....	11
ГЛАВА 2. АРХИТЕКТУРА ПРОЕКТА	13
2.1. Рассортировать по секциям	13
2.1.1. Реализация Dataset в morpheus.js	13
2.1.2. Стандартный класс ExpressionSet	13
2.1.3. opencpu.js	13
2.1.4. Protocol Buffers	13
2.1.5. Схема взаимодействия клиент-сервер	13
2.1.6. Загрузка и разбор данных из GEO	13
2.1.7. Загрузка данных в phantasus	13
2.1.8. Аннотации строк и столбцов матрицы	14

2.1.9. Git-репозиторий phantasmus	14
2.1.10. Запуск phantasmus локально и на сервере	14
Выводы к главе 2	14
ГЛАВА 3. РЕАЛИЗАЦИЯ	15
Выводы к главе 3	15
ЗАКЛЮЧЕНИЕ	16

ВВЕДЕНИЕ

ГЛАВА 1. ОБЗОР ПРЕДМЕТНОЙ ОБЛАСТИ

1.1. Биоинформатика

Биоинформатика — наука, объединяющая в себе методы прикладной математики, статистики, информатики для создания новых методов и алгоритмов для анализа разного рода биологических данных.

Биоинформатика занимается биохимией, биофизикой, экологией и многими другими областями биологии. Однако в данной работе фокус направлен на геномную биоинформатику и на конкретную ее задачу — анализ экспрессии генов.

1.1.1. Анализ экспрессии генов

Экспрессия генов — процесс преобразования наследственной информации от гена (в виде последовательности нуклеотидов ДНК) в функциональный продукт (РНК или белок).

Анализ экспрессии генов позволяет выяснить как ведет себя отдельный ген в разных условиях, тканях или организмах. Так, например, можно исследовать экспрессию вирусных белков или экспрессию онкогенов.

1.1.2. Используемые методы

Как было сказано ранее, биоинформатика использует в себе математику, информатику и статистику. Соответственно, задача анализа экспрессии генов сводится к исследованию путем статистических методов и алгоритмов числовой двумерной матрицы, как, например, в таблице ??.

Таблица 1.1 — Срез матрицы GSE14308. Строки матрицы соответствуют генам, столбцы — образцам.

	GSM357839	GSM357841	GSM357842	GSM357843	GSM357844
Rps29	16.32	16.30	16.25	16.32	16.30
Rpl13a	16.27	16.23	16.32	16.30	16.27
Rps3a1	16.23	16.19	16.30	16.25	16.25
Rpl38	16.21	16.25	16.27	16.27	16.21
Tmsb4x	16.30	16.32	16.23	16.21	16.32

На рисунке ?? можно увидеть визуализацию матрицы экспрессии в виде тепловой карты.

Также к основным методам анализа относятся:

- Иерархическая и вероятностная кластеризация;
- Дифференциальная экспрессия;
- Метод главных компонент и визуализация его результатов.

1.2. Существующие решения для анализа экспрессии генов

1.2.1. GENE-E

GENE-E - Платформа для анализа данных и визуального исследования данных, созданная на Java и R [?]. Содержит в себе множество полезных для исследования инструментов: тепловые карты, кластеризацию, фильтрацию, построение графиков и т.д. Позволяет исследовать любые данные в виде матрицы. К тому же, содержит дополнительные инструменты для геномных данных.

Недостатки:

- Чтобы использовать, необходимо устанавливать на свой компьютер;
- Поддержка данного приложения прекратилась в связи с созданием morpheus.js [?];
- Не имеет открытого исходного кода, а только API для взаимодействия и создания новых приложений на его основе.

1.2.2. morpheus.js

Morpheus.js - веб-приложение для визуализации и анализа матриц от создателя GENE-E. Создано уже на JavaScript и с открытым исходным кодом. Удобно для использования исследователями без навыков программирования и так же, как и GENE-E, применимо к любым матрицам.

Недостатки:

- Ограниченный набор функций, которых недостаточно для полноценного анализа;
- Для расширения биоинформатическими алгоритмами требуется реализовывать их заново на JavaScript.

1.2.3. R/Bioconductor

R - язык программирования для статистического анализа данных и работы с графикой [?].

Bioconductor - библиотека, содержащая в себе множество реализаций биоинформатических алгоритмов и методов обработки биологических данных на R. Она постоянно обновляется, пополняется новыми библиотеками, модерируется сообществом [?]. R и Bioconductor очень популярны в биоинформатической среде ввиду предоставляемых возможностей.

Однако для качественного и полноценного анализа с помощью этих инструментов, нужно иметь навыки программирования на R, что весьма неудобно для исследователей биологических специальностей.

1.3. Инструменты, которые могут быть применены

1.3.1. Язык R и библиотека Bioconductor

Как было сказано ранее, Bioconductor полон актуальными и широко используемыми биоинформатическими алгоритмами, в том числе и для анализа экспрессии генов. Соответственно, реализовывать их заново обычно нет необходимости и можно использовать их для достижения целей этой работы.

1.3.2. JavaScript

JavaScript — мультипарадигменный скриптов язык программирования, широко используемый для создания веб-приложений.

Node.js — [?]

1.3.3. R shiny

1.3.4. OpenCPU

Необходимые определения

HTTP

HTTP API

RPC

Веб-сервер

OpenCPU — система для встроенных научных вычислений и воспроизводимых исследований, предоставляющая HTTP API для взаимодействия с R-серверами [?]. Имеется также библиотека `оренспи.js` для интеграции JavaScript и R.

1.3.5. Gene Expression Omnibus

Gene Expression Omnibus (GEO) — международный публичный репозиторий, агрегирующий и распространяющий различные формы геномных данных от исследовательского сообщества [?].

В библиотеке Bioconductor есть R-пакет `GEOquery` для удобной загрузки данных из GEO.

1.3.6. Docker

1.3.7. JSON

1.3.8. Protocol Buffers

1.3.9. Apache

1.3.10. HTML

1.4. Постановка задачи

Рассмотрев существующие решения для анализа экспрессии генов и инструментов, которые могли бы пригодиться для будущих решений, можно сформулировать цель и основные задачи данной работы

1.4.1. Цель работы

Создать веб-приложение, интегрирующее существующие возможности веб-приложения `morpheus.js` и методы анализа, реализованные в Bioconductor.

1.4.2. Основные задачи

- а) Разработать способ взаимодействия между js-клиентом и R и встроить его в morpheus.js, чтобы избежать реализации с нуля уже существующих алгоритмов;
- б) Реализовать графический интерфейс в js-клиенте и серверную реализацию в R-пакете;
- в) Соединить все составляющие в одном веб-приложении phantasmus;
- г) Запустить веб-приложение в открытый доступ для исследователей.

1.4.3. Требования к веб-приложению phantasmus

Доступность

Необходимо, чтобы веб-приложение phantasmus было доступно для исследователей независимо от их местоположения и времени суток. Варианты действий:

- а) Сделать его доступным по определенному веб-адресу, и тогда пользователь сможет продолжать исследования из любой точки, где есть подключение к интернету;
- б) Предоставить возможность запускать приложение локально, например, с помощью Docker или внутри R.

Возможность дальнейшего расширения функционала веб-приложения

Как уже было сказано выше, библиотека Bioconductor постоянно обновляется и пополняется новыми алгоритмами, а исследователи находят новые методы для анализа экспрессии генов, так что необходимо не только реализовать дополнительные методы, но и отладить и описать алгоритм действий для добавления новых.

Выводы по главе 1

В данной главе была кратко описана предметная область и необходимые биоинформатические определения, рассмотрены существующие решения и

инструменты, которые могли бы быть применены для разработки новых решений. Исходя из обзора, была сформулирована цель работы и требования к результату:

- а) доступность;
- б) возможность дальнейшего расширения функционала.

ГЛАВА 2. АРХИТЕКТУРА ПРОЕКТА

В этой главе будут подробно рассмотрены элементы проекта, их взаимосвязь и ключевые для архитектуры выдержки из исходного кода.

2.1. Рассортировать по секциям

2.1.1. Реализация Dataset в morpheus.js

В исходном коде morpheus.js имеется класс для работы с данными, позволяющий рассматривать сечения, работать с аннотациями, в целом, осуществлять любое взаимодействие с имеющимися данными. Код представлен на листинге 2.1.

Листинг 2.1 — класс Dataset

2.1.2. Стандартный класс ExpressionSet

2.1.3. opencpu.js

2.1.4. Protocol Buffers

На стороне клиента

На стороне сервера

2.1.5. Схема взаимодействия клиент-сервер

2.1.6. Загрузка и разбор данных из GEO

2.1.7. Загрузка данных в phantasus

Из файла

- а) My computer;
- б) URL;
- в) Dropbox.

Из GEO

2.1.8. Аннотации строк и столбцов матрицы

2.1.9. Git-репозиторий phantasm

2.1.10. Запуск phantasm локально и на сервере

Выводы к главе 2

ГЛАВА 3. РЕАЛИЗАЦИЯ

Выводы к главе 3

ЗАКЛЮЧЕНИЕ