Министерство образования и науки Российской Федерации
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«САНКТ-ПЕТЕРБУРГСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ»

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

АВТОМАТИЧЕСКОЕ ВЫДЕЛЕНИЕ ФУНКЦИОНАЛЬНЫХ НАБОРОВ ГЕНОВ В БАЗЕ GEO С ПОМОЩЬЮ ДИФФЕРЕНЦИАЛЬНОЙ ЭКСПРЕССИИ

Автор: Беляева Екатерина Сергеевна _____
Направление подготовки (специальность): 09.03.02 Информационные системы и технологии
Квалификация: Бакалавр
Руководитель: Сергушичев А.А., к.т.н. _____

К защите допустить
Зав. кафедрой Лисицына Л.С., д.т.н., проф. _____

«__» _____ 20__ г.

Санкт-Петербург, 2018 г.

Студент Беляева Е.С. **Группа** Р3420 **Кафедра** КОТ **Факультет** ПИиКТ

Направленность (профиль), специализация Автоматизация и управление в образовательных системах

ВКР принята «___» _____ 20___ г.

Оригинальность ВКР _____%

ВКР выполнена с оценкой _____

Дата защиты «19» июня 2018 г.

Секретарь ГЭК *Бутько Е.Ф.*_____
Листов хранения _____

Демонстрационных материалов/Чертежей хранения _____

Министерство образования и науки Российской Федерации

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«САНКТ-ПЕТЕРБУРГСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ»

УТВЕРЖДАЮ

Зав. кафедрой КОТ

Лисицына Л.С. _____

«___» _____ 20___ г.

ЗАДАНИЕ

НА ВЫПУСКНУЮ КВАЛИФИКАЦИОННУЮ РАБОТУ

Студент Беляева Е.С. **Группа** Р3420 **Кафедра** КОТ **Факультет** ПИиКТ
Руководитель Сергушичев А.А., к.т.н., доцент кафедры КТ

1 Наименование темы: Автоматическое выделение функциональных наборов генов в базе GEO с помощью дифференциальной экспрессии

Направление подготовки (специальность): 09.03.02 Информационные системы и технологии
Направленность (профиль): Автоматизация и управление в образовательных системах
Квалификация: Бакалавр

2 Срок сдачи студентом законченной работы: «31» мая 2018 г.

3 Техническое задание и исходные данные к работе.

В рамках выпускной квалификационной работы основной задачей является построение функциональных наборов генов из базы GEO с помощью анализа дифференциальной экспрессии, а также обеспечение интерфейса доступа к ним

4 Содержание выпускной квалификационной работы (перечень подлежащих разработке вопросов)

а) Обзор предметной области.
б) Анализ существующих сервисов.
в) Загрузка данных и их обработка.
г) Построение модулей генов при помощи анализа дифференциальной экспрессии.
д) Обеспечение интерфейса доступа к модулям генов.
е) Анализ полученных результатов.

5 Перечень графического материала (с указанием обязательного материала)

Не предусмотрено

6 Исходные материалы и пособия

а) Bioconductor. Bioconductor is an open source, open development software project to provide tools for the analysis and comprehension of high-thoughput genomic data. [Электронный ресурс]. URL: https://www.bioconductor.org/;
б) Docker. Docker is the software container platform. [Электронный ресурс]. URL: https://www.docker.com/;

в) Терри А.Браун Геномы / Терри А.Браун. - М.-Ижевск: Институт компьютерных исследований, 2011. - 921 с.

7 Дата выдачи задания: «01» сентября 2017 г.

Руководитель ВКР _____

Задание принял к исполнению _____ «01» сентября 2017 г.

Министерство образования и науки Российской Федерации

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«САНКТ-ПЕТЕРБУРГСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ»

АННОТАЦИЯ
ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЫ

Студент: Беляева Екатерина Сергеевна

Наименование темы работы: Автоматическое выделение функциональных наборов генов в базе GEO с помощью дифференциальной экспрессии

Наименование организации, где выполнена работа: Университет ИТМО

ХАРАКТЕРИСТИКА ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЫ

1 Цель исследования: Получить новую базу функциональных наборов генов.

2 Задачи, решаемые в работе:

а) изучение существующих сервисов для работы с данными экспрессии генов;
б) построение новых функциональных наборов генов из базы GEO при помощи анализа дифференциальной экспрессии;
в) анализ полученных результатов;
г) обеспечение интерфейса доступа к модулям генов.

3 Число источников, использованных при составлении обзора: 3

4 Полное число источников, использованных в работе: 10

5 В том числе источников по годам

Отечественных			Иностранных		
Последние 5 лет	От 5 до 10 лет	Более 10 лет	Последние 5 лет	От 5 до 10 лет	Более 10 лет
1	0	0	8	0	1

6 Использование информационных ресурсов Internet: да, число ресурсов: 5

7 Использование современных пакетов компьютерных программ и технологий:

Пакеты компьютерных программ и технологий	**Параграф работы**
R (библиотеки GEOquery, limma, dplyr, stringr, data.table, magrittr, tibble)	Глава 2, Глава 3
Python 2,3 (библиотеки pandas, subprocess, re, csv)	Глава 2, Глава 4
Bash	Глава 2
Kotlin	Глава 2, Глава 4
Django	Глава 2, Глава 4
Java Script	Глава 2, Глава 4
React	Глава 2, Глава 4
HTML	Глава 2, Глава 4
CSS	Глава 2, Глава 4
Docker	Глава 2, Глава 4

8 Краткая характеристика полученных результатов: По итогу работы был реализован сервис с возможностью предоставления доступа к модулям генов конечному потребителю.

9 Гранты, полученные при выполнении работы: Грантов или других форм государственной поддержи и субсидирования в процессе работы не предусматривалось.

10 Наличие публикаций и выступлений на конференциях по теме работы: Публикаций и выступлений на конференциях не было.

Выпускник: Беляева Е.С. _____

Руководитель: Сергушичев А.А. _____

«__» _____ 20__ г.

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Жизнь, какой мы ее видим, создается геномами мириад организмов, с которыми мы делим нашу планету. Каждый из этих организмов обладает геномом, содержащем биологическую информацию, необходимую для построения и поддержания организма живущего в настоящий момент времени представителя данного вида.[1] Биоинформатика – наука, стоящая на границе биологии и информатики, активно развивается в течении последних десятилетий. Открываются новые направления исследований, появляется всё больше данных, в том числе и открытых.

Одним из направлений биоинформатики являются исследования, связанные с экспрессией генов, т.е. с процессом преобразования последовательности гена в функциональный продукт (обычно белок). Существует множество связанных с этим данных. Самой известной и часто используемой базой данных экспрессии генов является база данных Gene Expression Omnibus (GEO). Она открытая, и именно она и будет использована в дальнейшей работе.

Есть множество сервисов, работающих с данными экспрессии генов. В этой работе был рассмотрен и расширен сервис GeneQuery, который является поисковым порталом, использующим данные из базы GEO. Сервис плохо работал с маленькими по числу образцов датасетами, следовательно существовала проблема изменения принципа обработки данных. Так, целью данной бакалаврской работы являлось построение функциональных наборов генов из базы GEO с помощью дифференциальной экспрессии. Сервис GeneQuery был расширен для поддержки новых данных. Ожидается, что используя новую базу функциональных групп генов может быть выявлена потенциально значимая биологическая информация, а сам сервис может быть использован исследователями в области биоинформатики.

ГЛАВА 1. ОБЗОР ПРЕДМЕТНОЙ ОБЛАСТИ

1.1. Основные понятия и определения

Биоинформатика – это междисциплинарная область, работающая с биологическими данными. Она сочетает в себе информатику, биологию, математику, статистику и ещё множество областей. Также широк и круг исследований биоинформатики, однако данная работа будет сосредоточена на таком направлении, как экспрессия генов.

Ген – единица наследственности живых организмов.

Экспрессия гена – процесс преобразования последовательности гена в функциональный продукт (обычно белок). Она может быть измерена количественно.

Дифференциальная экспрессия гена – изменение уровня экспрессии гена в зависимости от биологического состояния.

Функциональная группа генов – отвечает за конкретное биологическое явление или несколько явлений.

Для примера можно рассмотреть гликолиз – процесс расщепления глюкозы в клетках, сопровождающийся синтезом АТФ. Для протекания этого процесса необходимы ферменты, т.е. ускорители химических реакций. В гликолизе участвуют ферменты гексокиназа, глюкозофосфатизомераза, енолаза и другие. Их структуры определяются генами HK1, PGI1, ENO1 и т.д.[2] Все эти гены являются функциональной группой генов для процесса гликолиз.

1.2. База данных GEO

Биоинформатика обладает большим объемом информации, который нужно где-то хранить. Для этой цели существуют множество баз данных. Накопление данных о генах, проявивших себя в биологических экспериментах, т.е. функциональных групп генов, является важной задачей потому, что эта информация может быть полезна в будущем. Самым известным открытым репозиторием для хранения данных экспрессии генов является *Gene Expression Omnibus*. Или же сокращенно *GEO*.

Проект GEO был инициирован в ответ на растущую потребность в общедоступном хранилище данных экспрессии генов.[3] Эксперименты, хранящиеся в репозитории, имеют определенную структуру. Для каждого эксперимента представлены:

— GSM (Geo SaMple) – *образцы* или *сэмплы* – эти данные содержат информацию об организмах, участвовавших в эксперименте, об условиях эксперимента.
— GSE (Geo SEries) – *серия* – объединяет в себе несколько образцов и хранит данные о числовых значениях экспрессии генов каждого образца. Как правило это 20000 генов на каждый обзазец.
— GPL (Geo PLatform) – *платформа*, на которой вычислялась экспрессия. Важным компонентом, хранящимся здесь, является информация для каждого гена о соответствии его entrez к symbol.

Количество данных GEO огромно и постоянно растет, в следствии чего база представляет исключительный интерес в биоинформатическом сообществе. Как было отмечено ранее, данная работа посвящена обработке данных именно из GEO.

1.3. Сервис GeneQuery

Проект GeneQuery[4] - это поисковый портал, основанный на автоматическом выделении наборов генов с помощью кластеризации. Работает с данными экспрессии генов из базы данных GEO для организмов Homo Sapience, Mus Musculus и Rattus Norvegicus, т.е. с человеком, мышью и крысой.

GeneQuery имеит такие достоинства, как высокая скорость обновления данных (новые данные из GEO можно портировать в GeneQuery), быстрая скорость выдачи результата для пользователя, простота использования. Сервис достаточно хорошо работает для средних по размеру датасетов (по числу образцов), но не подходит для маленьких. Для последних можно использовать анализ дифференциальной экспрессии, что и сделано в альтернативной версии GeneQuery - в GeneQueryDE.

1.4. Цель и актуальность работы

Целью данной работы являлось построение наборов генов по экспериментам из базы GEO с помощью анализа дифференциальной экспрессии и обеспечение интерфейса доступа к ним.

Портал GeneQuery расширен для поддержки новых данных до версии GeneQueryDE. Сервис может быть использован исследователями в области биоинформатики. Ожидается, что используя новую базу функциональных групп генов может быть выявлена потенциально значимая биологическая информация.

Выводы по главе 1

В данной главе был произведен обзор предметной области. А именно, рассмотрены основные понятия, являющиеся базовыми для понимания дальнейших действий. Рассмотрены сервисы, на которых строится данная работа, а также обозначены её цель и значимость.

ГЛАВА 2. СХЕМА СЕРВИСА

Взаимодействие элементов системы представлено на рисунке 1.

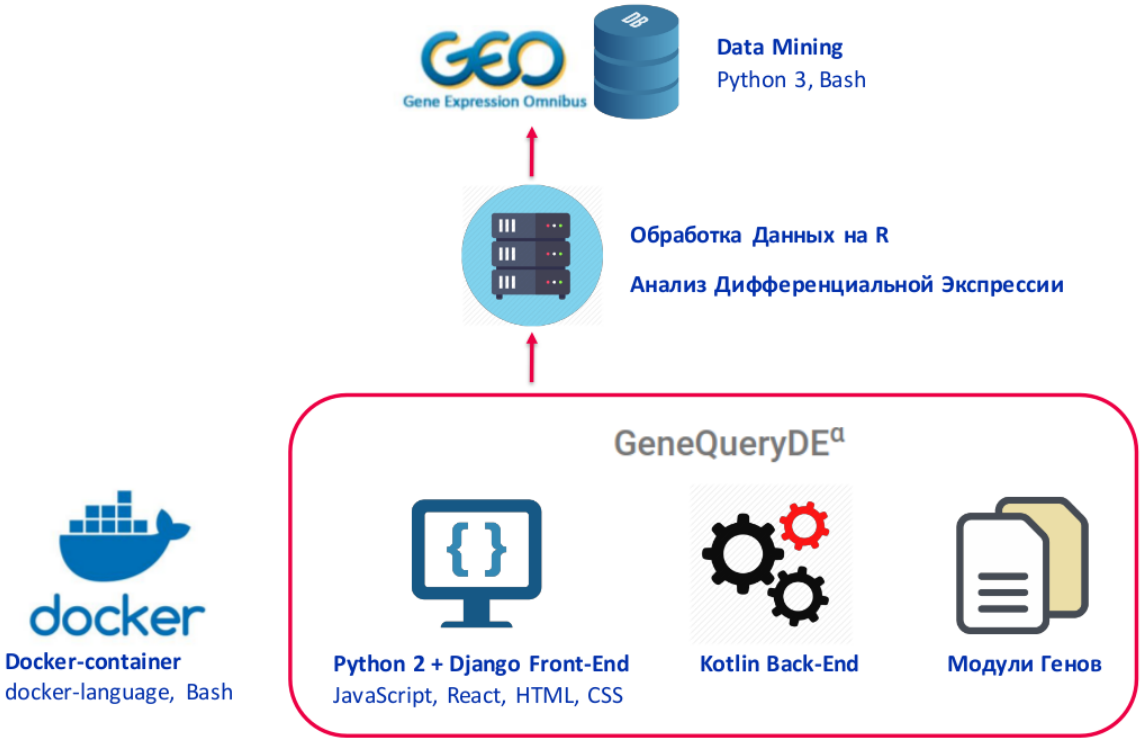

Рисунок 1 – Схема сервиса.

С начала данные были загружены из базы GEO на сервер. Для этого использовались языки Python 3 и Bash. Далее на сервере происходила их обработка, и при помощи анализа дифференциальной экспрессии строились модули генов. Для этого использовался язык R. Bioconductor – это проект на R, содержащий множество библиотек для анализа геномных данных.[5]. В том числе имеет библиотеку для взаимодействия с базой данных GEO. Так, работая на R, были использованы пакеты GEOquery и limma.

После того, как модули генов были построены, сервис GeneQuery был расширен для поддержки новых данных до версии GeneQueryDE. Сам сервис состоит из двух репозиториев - front-end части на Python 2 + Django и back-end части на Kotlin. После, GeneQueryDE с новыми данными был упакован в docker-контейнер и запущен. Сейчас сервис доступен по адресу *http://genome.ifmo.ru/genequery-de*.

Выводы по главе 2

В данной главе была представлена схема работы системы и описаны основные использованные технологии.

ГЛАВА 3. ПОСТРОЕНИЕ МОДУЛЕЙ ГЕНОВ

3.1. Загрузка данных

База данных GEO имеет сложную структуру. Например, датасет GSE40554-GPL4125 находится по адресу *https://ftp.ncbi.nlm.nih.gov/geo/series/GSE40nnn/GSE40554/matrix/GSE40554-GPL4125_series_matrix.txt.gz*, а аннотация к нему лежит на *https://ftp.ncbi.nlm.nih.gov/geo/platforms/GPL4nnn/GPL4125/annot/GPL4125.annot.gz*. При скачивании данных, структура GEO была сохранена.

```
import subprocess
import csv

def download_data(file_name, kind_data):
    with open(file_name,'r') as tsvin:
        tsvin = csv.reader(tsvin, delimiter='\t')
        indexes = []

        for (i, row) in enumerate(tsvin):
            if i == 0:
                if kind_data == 'series':
                    indexes = [row.index(j) for j in ['Series', 'Series_url']]
                else:
                    indexes = [row.index(j) for j in ['Platforms', 'Platforms_url
                        ']]
            else:
                try:
                    dif_string = 'https://ftp.ncbi.nlm.nih.gov/'
                    path = row[indexes[1]].replace(dif_string, '')
                    path = path[:(path.rindex('/')+1)]
                    directory = "../" + path

                    subprocess.call(["mkdir", "-p", directory])
                    subprocess.call(["wget", "-c", "-P", directory, row[indexes
                        [1]]])

                except:
                    pass
```

Листинг 1 – Загрузка данных из GEO.

Основная функция загрузки приведена на листинге 1. Всего было загружено 74177 датасетов типа GSE (с экспрессией генов) и 790 GPL (аннотаций) для видов Homo Sapience, Mus Musculus и Rattus Norvegicus. Далее GSE были сопоставлены с аннотациями и выяснилось, что 29098 датасетов пригодно для построения таблиц дифференциальной экспрессии.

3.2. Автоматическая обработка

После получения таблицы с матчингом GSE с GPL, для каждого эксперимента надо было построить модули генов с применением анализа дифференциальной экспрессии. Для этого, для каждого датасета были выполнены действия:

— Выявлены уникальные условия каждого эксперимента.
— Соотнесены гены с аннотациями к ним.
— Нормализованны данные.
— Построены пары условий.
— Построены таблицы дифференциальной экспрессии.

3.2.1. Выявление уникальных условий каждого эксперимента

GSE датасет содержит таблицу с указанием информации по каждому образцу, участвовавшему в эксперименте. В том числе, содержатся условия эксперимента для каждого образца. Эти данные записаны в столбцах «characterics», которых может быть от 1 до n. Требовалось автоматически выделять биологические состояния.

Поиск нужных для парсинга столбцов на листинге 2.

```
getCharacteristicsColumns <- function(gse) {
  col <- colnames(pData(gse))
  characteristics <- c()
  for (ch in col) {
    if (grepl("characteristics", ch)) characteristics <- c(characteristics, ch)
  }
  return(characteristics)
}
```

Листинг 2 – Поиск столбцов «characterics».

Как можно видеть на листинге 3, возможна ситуация, когда уникальные состояния выделить не удавалось. Это могло быть связано с пропусками данных в столбцах «characterics» или же с тем, что условия не несли в себе универсального значения. Например, в датасете, состоящем из восьми образцов, в одной из колонок «characterics» могли быть написаны цифры от 1 до 8, соответствующие просто номерам образцов. Такие случаи игнорировались и не участвовали в дальнейшем анализе.

```
characteristics <- getCharacteristicsColumns(gse)
conditionLists <- list()

...

if (length(characteristics) > 0) {
    message("There are characteristics columns in the samples table.")
    conStructure <- getConditionsFromCharacteristics(gse, characteristics)
    conditionLists <- conStructure$conditionsList
    explanatoryTable <- conStructure$explanatoryTable
  }

  if (length(conditionLists) > 0 && length(unique(unlist(conditionLists))) > 1) {
    pData(gse)$condition <- fillGseConditionColumn(conditionLists)
    ...

  } else
    message("Characteristics columns in the samples table don't exist or were
        unhelpful.")
```

Листинг 3 – Выделение условий.

3.2.2. Соотнесение генов и аннотаций

Требовалось соотнести entrez генов из таблицы с экспрессией с их symbol из таблицы с аннотацией. При построении модулей дифференциальной экспрессии колонка с названиями генов также была сохранена в таблице.

```
collapseData <- function(gse, gpl, FUN=median) {
  ranks <- apply(exprs(gse), 1, FUN)
  ranks <- data.frame(r = ranks, i = seq_along(ranks))
  table <- inner_join(rownames_to_column(gpl), rownames_to_column(ranks),
            by="rowname") %>%
            mutate(j = seq_along(symbol))
  t <- table[order(table$r, decreasing=T), ][1:7000,]
  keep <- t$i
  res <- gse[keep, ]
  rownames(res) <- table$ENTREZ_GENE_ID[t$j]
  fData(res)$symbol <- t$symbol
  return(res)
}

...

gpl <- createGenesSymbolsTable(gpl)
es <- collapseData(gse, gpl)
```

Листинг 4 – Выделение 7000 генов с максимальной экспрессией.

На листинге 4 происходит матчинг entrez генов с их symbol, выделяются 7000 генов с наибольшей экспрессией. На основе этого набора генов в дальнейшем строются модули с дифференциальной экспрессией.

3.2.3. Нормализация данных

В процессе обработки данных были совершены следующие действия:

— Убраны дублирующиеся данные.
— Убраны данные с пропусками.
— Отобраны 7000 генов с самой высокой экспрессией (листинг 4).
— Логарифмизация данных при большом размахе значений экспрессии (листинг 5).

```
if (max(exprs(es)) - min(exprs(es)) > 100)
        exprs(es) <- normalizeBetweenArrays(log2(exprs(es)+1), method="quantile")
```

Листинг 5 – Логарифмизация данных.

Как правило, после удаления дублирующихся данных и данных с пропусками в рассмотрении оставался набор, состоящий примерно из 20000 генов.

3.2.4. Построение пар условий

Так как дифференциальная экспрессия – это изменение активности гена в зависимости от конкретного биологического состояния, то на данном этапе требовалось составить пары условий с различием в одно состояние. Для этого обрабатывался столбец *pData(gse)$condition*, получение которого показано на листинге 3.

3.2.5. Построение таблиц дифференциальной экспрессии

Для каждой пары условий, полученной по результатам предыдущего пункта, требовалось построить таблицу с дифференциальной экспрессией. Построение показано на листинге 6.

Пример таблицы приведен на рисунке 2. Модуль содержит 7000 генов и включает в себя информацию о entrez гена, его названии, средней экспрессии и различные статистические метрики.

Модули получилось построить для 16246 экспериментов.

3.3. Верхняя и нижняя регуляции генов

Верхняя и нижняя регуляция генов – это сравнение их активностей относительно друг друга. На предыдущем этапе были постоены таблицы генов при противопоставлении двух разных биологических состояний. Так, на данном этапе для каждого из модулей дифференциальной экспрессии генов требовалось

```
fitLinearModel <- function(fit, conditions, design, deSize) {
  deList <- list()
  for (i in 1:length(conditions)) {

    contrasts <- makeContrasts2(
                 c("condition", conditions[[i]]$firstCon, conditions[[i]]$secondCon
                     ),
                 levels=design)

    fit2 <- contrasts.fit(fit, contrasts)

    df.residual <- unique(fit2$df.residual)
    if ((length(df.residual) == 1 && df.residual[1] != 0) ||
        (length(df.residual) != 1)) {

      fit2 <- eBayes(fit2)

      deList[[i]] <- data.table(
        topTable(fit2, adjust.method="BH", number=deSize, sort.by = "B"),
        keep.rownames = T)
    }
  }
  return(deList)
}
```

Листинг 6 – Построение модулей генов при помощи дифференциальной экспрессии.

	A	B	C	D	E	F	G	H
1	rn	symbol	logFC	AveExpr	t	P.Value	adj.P.Val	B
2	20715	Serpina3g	6.030547534	10.6892139601	63.5082547013	5.52118363962044E-16	3.86482854773431E-12	25.902659987
3	17329	Cxcl9	6.3356914437	11.1081628067	50.3345597005	7.89521658726984E-15	2.76332580554445E-11	23.9421695132
4	433470	AA467197	4.8914732002	10.1720346836	48.1240467197	1.31922974800399E-14	3.07820274534264E-11	23.5316561252
5	16153	Il10	4.2295908369	10.1817500872	45.4194949439	2.55467298396242E-14	4.47067772193424E-11	22.9893089008
6	21939	Cd40	4.9775575043	11.5357658018	43.8679282341	3.79952707161957E-14	5.3193379002674E-11	22.6563788376
7	238393	Serpina3f	4.4428201703	9.8144074483	42.9611188811	4.82280357789186E-14	5.40983082981626E-11	22.4538784664
8	667373	Ifit1bl1	6.4006078219	10.8477655629	42.5309949304	5.40983082981626E-14	5.40983082981626E-11	22.3556956694
9	215900	Fam26f	4.9573906927	10.6700218457	38.0833579053	1.90701284723663E-13	1.66863624133205E-10	21.2525205748
10	20293	Ccl12	4.8845394113	10.2466374648	33.9997416566	6.93922387707301E-13	5.39717412661234E-10	20.0762550528
11	57444	Isg20	4.4238649325	9.8502322878	33.6162522149	7.89498185435406E-13	5.52648729804784E-10	19.9564774078
12	20306	Ccl7	5.0046765276	10.3295731048	33.1902740896	9.12726442872727E-13	5.80825918191736E-10	19.8213859834
13	14469	Gbp2	6.1048408097	12.3552471869	30.4561553354	2.42406699001782E-12	1.12203333713762E-09	18.8995639495
14	20440	St6gal1	-3.0680562242	9.9700985417	-30.3521433526	2.52004868904069E-12	1.12203333713762E-09	18.8625045566
15	434341	Nlrc5	3.5299944374	10.1806904955	30.2039436948	2.66400828460823E-12	1.12203333713762E-09	18.8094339419
16	20296	Ccl2	3.2613845656	9.3364759317	30.1773683194	2.69075540828832E-12	1.12203333713762E-09	18.7998838549
17	384009	Glipr2	3.4399433181	9.3336980256	30.1570985832	2.71135219609879E-12	1.12203333713762E-09	18.7925929003
18	55932	Gbp3	5.7307955453	11.0316049778	30.139529113	2.72934372101617E-12	1.12203333713762E-09	18.7862684181
19	623121	Pydc4	3.4673625155	9.1939253204	29.9923720691	2.88522858121102E-12	1.12203333713762E-09	18.7331204748
20	107607	Nod1	3.6845924408	10.2540842923	29.433783429	3.57115125319917E-12	1.31434428741991E-09	18.5284760793
21	14190	Fgl2	4.3861676633	9.9567512421	29.3036076285	3.75526939262832E-12	1.31434428741991E-09	18.4801135337
22	17858	Mx2	4.9663078784	10.6244240359	28.4180215472	5.31776893282039E-12	1.77258964427346E-09	18.1441434238
23	15945	Cxcl10	5.0658835144	10.2232050734	28.1323356722	5.9627956451799E-12	1.89725315982997E-09	18.0331002325
24	54199	Ccrl2	3.3698242532	9.651375616	27.9820867281	6.33573059337887E-12	1.89888106503496E-09	17.974164357
25	81913	Bambi-ps1	3.292733862	9.3821770387	27.8279828643	6.74470365158551E-12	1.89888106503496E-09	17.9133279097
26	74481	Batf2	3.7166394137	9.5759254503	27.8145391522	6.78171808941057E-12	1.89888106503496E-09	17.9080018755
27	16365	Acod1	5.1400474315	10.5609573344	27.21930329	8.66390684701144E-12	2.33259030496462E-09	17.6691132718
28	22035	Tnfsf10	3.7883540799	9.4726720517	26.8191061274	1.02449835098398E-11	2.65610683588439E-09	17.5050462442

Рисунок 2 – Таблица дифференциальной экспрессии генов.

построить дескриптивную выборку. По максимальному значению t-статистики отбиралось 200 генов, что соответствует верхней регуляции и по минимально-

му значению t-статистики отбиралось 200 генов, соответствующих нижней регуляции. Полученные модули сохранялись. Также на их основе строились *.gmt* файлы для каждого вида (человек, мышь, крыса). Файлы содержали гены из экспериментов. А именно, для каждого эксперимента было представленно название, универсум (7000 генов), название модулей и соответствующие выборки по 200 генов.

По итогам выборак по 200 генов, всего было получено 851396 новых модулей, которые затем были встроены в обработку сервиса GeneQueryDE.

Выводы по главе 3

В данной главе был описан основной этап работы, который заключался в загрузке данных из базы GEO и построении модулей генов. Приведены цифры обработки данных и итоговые результаты.

ГЛАВА 4. ОБЕСПЕЧЕНИЕ ИНТЕРФЕЙСА ДОСТУПА К ДАННЫМ

После получения модулей генов и *.gmt* файлов требовалось расширить сервис GeneQuery для возможности работы с новыми данными. Так была создана версия проекта GeneQueryDE.

4.1. Метод поиска модулей генов

GeneQueryDE со стороны пользователя принимает название организма и некий набор генов. Затем на основании статистической значимости выдаются подобранные модули. Для этого отдельно оценивается значимость каждого модуля, используя тест Фишера. При этом обозначаются две гипотезы:

— *нулевая* – связи между двумя наборами генов нет, пересечение случайно.

— *альтернативная* – пересечение получено не случайно, наблюдается статистически значимая связь, гены связанны одним процессом.

Таблица 1 – Тест Фишера

	в запросе	вне запроса	всего
в модуле	a	b	a + b
не в модуле	c	d	c + d
всего	a + c	b + d	n

, где *a, b, c, d* - количества генов в соответствующих категориях;

n - размер универсума (= 7000 генов)

$$p = \binom{a+b}{a}\binom{c+d}{c} / \binom{n}{a+c} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!a!b!c!d!} \qquad (1)$$

Используя формулу 1 производится расчет таблицы 1. Чем меньше полученное *p-value*, тем значимее модуль. Таким образом, число выражает нашу «неуверенность» в связи модуля и запроса.

Далее происходит корректировка на множественные сравнения. Для этого применяется поправка Бонферрони[6] – полученное *p-value* умножается на количество модулей, т.е. на число произведенных сравнений. Полученное значение после поправки, *adj.p-value*, равно вероятности отклонить хотя бы одну нулевую гипотезу из всех проверенных, т.е. вероятности совершить хотя бы одну ошибку первого рода. Если *adj.p-value* ≤ 0.01, то модуль признается статистически значимым и выводится в качестве результата запроса.

4.2. Представление результатов пользователю

Пользователь выбирает организм и вводит соответствующий ему набор генов, как показано на рисунке 3. Система выдает ответ в виде статистически значимых модулей (рисунок 4), в которых найдены совпадающие гены. Список отсортирован в порядке возрастания логарифма *adj.p-value* ≤ 0.01. Как можно видеть, информация, которую получает пользователь по каждому модулю, также включает в себя название эксперимента, название модуля, число перекрытий генов, ссылку на эксперимент и возможность загрузки таблицы с дифференциальной экспрессией данного модуля.

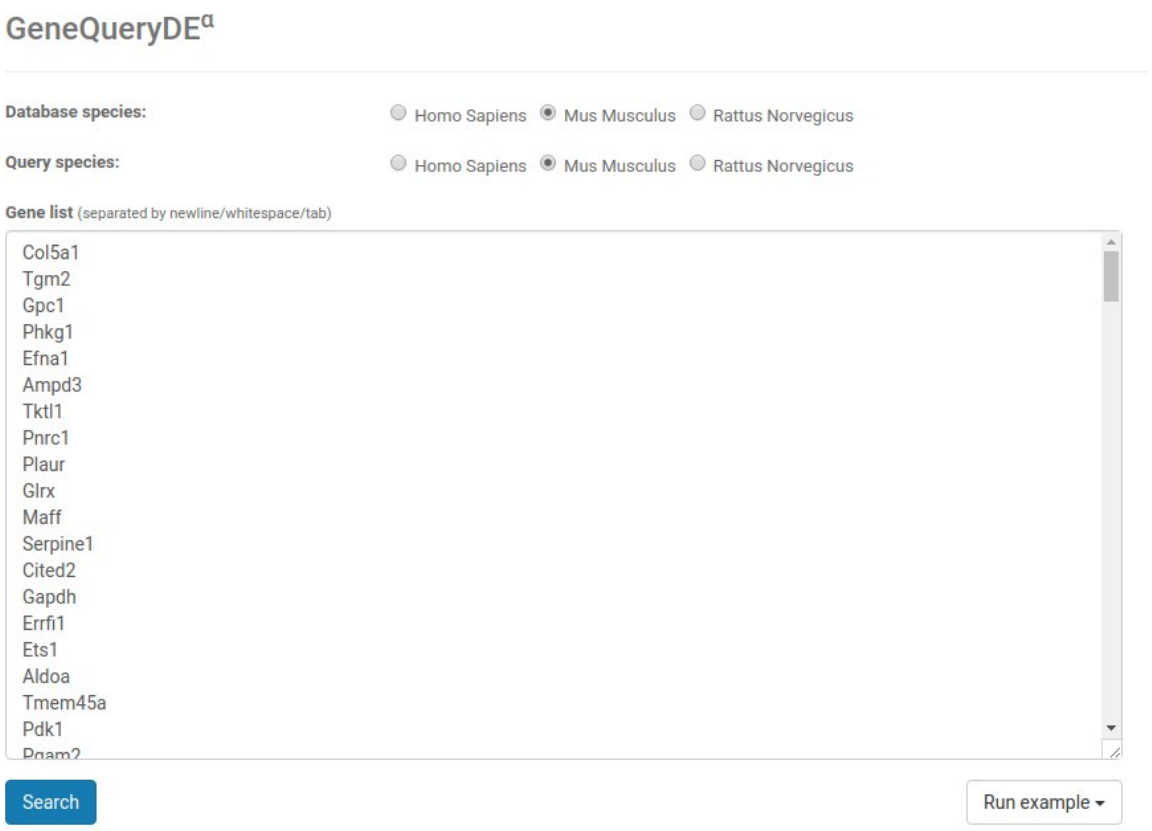

Рисунок 3 – Запрос.

Название модуля включает в себя два биологических состояния, на основании которых была вычисленна дифференциальная экспрессия, и условия, которые не менялись.

Например, *up in* **‘oxygen.3%.O2’** *compared to* **‘oxygen.21%.O2’** *on ‘tissue.fetal.cortex_Cortex.primary.culture_13.days’ background* – ответ на гипоксию. Здесь гены при **‘oxygen.3%.O2’** реагируют против **‘oxygen.21%.O2’** на основании условий, присутствующих в обоих экспериментах.

#	Experiment title	Module	$log_{10}(adj.p_{value})$	Overlap	GSE	Dif exprs
1	Expression data from mouse B16-F10 cells exposed to hypoxic conditions in vitro.	up in **'condition.exposed.to.experimental.hypoxia.(1%.oxygen).in.incubator'** compared to **'condition.cultured.in.control.conditions.(21%.oxygen)'**	-45.56	57/200	GSE33607	
2	Whole transcript expression profiling of short-term and long-term hypoxic neural stem cells (NSCs)	up in **'oxygen.3%.O2'** compared to **'oxygen.21%.O2'** on 'tissue.fetal.cortex_Cortex.primary.culture_13.days' background	-38.96	46/200	GSE80070	
3	HIF1a-dependent glycolytic pathway orchestrates a metabolic checkpoint for the differentiation of TH17 and Treg cells	up in **'genotype.variation.wild.type'** compared to **'genotype.variation.HIF1a.deficient'**	-36.81	42/200	GSE29765	
4	Expression data from E2f7/E2f8/E2f3a null placentas and embryos	up in **'genotype.Cyp19Cre.E2f7F.F.E2f8F.F'** compared to **'genotype.E2f3a+/+;E2f7+/+;E2f8+/+'** on 'tissue.Embryo' background	-32.71	42/200	GSE30488	
5	HL-1 cardiomyocyte response to hypoxia	up in **'culture.Hypoxia'** compared to **'culture.Normoxia'**	-31.51	40/200	GSE27975	
6	Whole gene expression data from osteocyte-like cell line MLO-Y4 under large gradient high magnetic field (LG-HMF)	up in **'treatment.LG-HMF.2-g'** compared to **'treatment.LG-HMF.μ-g'**	-30.70	42/200	GSE62128	
7	Expression data from mouse B16-F10 cells exposed to hypoxic conditions in vitro.	up in **'condition.exposed.to.100.μM.CoCl2'** compared to **'condition.cultured.in.control.conditions.(21%.oxygen)'**	-30.35	46/200	GSE33607	

Рисунок 4 – Результат запроса.

4.3. Контейнеризация

В соответствии с документацией docker[7] был создан Dockerfile для построения образа и впоследующем контейнера с сервисом. Используя данную технологию, проект GeneQueryDE, состоящий из front-end и back-end частей был упакован в docker-контейнер с доступом к данным с сервера – к модулям генов и *.gmt* файлам. Контейнер запущен, т.ч. сервис доступен по адресу *http://genome.ifmo.ru/genequery-de*.

Выводы по главе 4

В данной главе был описан интерфейс сервиса GeneQueryDE, способ выдачи результата и пример запроса-ответа. Также сказано о возможности доступа к сервису посредством интернета.

ГЛАВА 5. СРАВНЕНИЕ И АНАЛИЗ РЕЗУЛЬТАТА

5.1. Сравнение GeneQuery с GeneQueryDE по числу данных

На рисунке 5 представленно сравнение сервисов по количеству пар GSE+GPL, для которых в последствии были построены модули генов. Как можно увидеть, помимо совпадающих экспериментов, проанализированных в обоих сервисах, также много уникальных GSE, обработанных только в одном из проектов. Причем в GeneQueryDE обработано больше. Если для вида Rattus Norvegicus количество уникальных GSE примерно равно в обоих проектах, то, например, для Mus Musculus, сервис GeneQueryDE обрабатывает в 3.15 раза больше данных, чем версия GeneQuery.

Рисунок 5 – Сравнение по парам GSE+GPL.

В таблице 2 представлены различия в итоговом числе модулей, полученных для двух сервисов. Видно, что сервис GeneQueryDE работает с гораздо большим количеством данных, чем альтернативная версия.

Таблица 2 – Сравнение по модулям генов

	Модули экспрессии генов	
	GeneQuery	GeneQueryDE
Homo Sapience	117497	620216
Mus Musculus	82371	213712
Rattus Norvegicus	13560	17468

5.2. Сравнение сервисов

В данном разделе приведено сравнение существующих проектов с GeneQueryDE. А именно, сервисов GeneQuery, GEOracle[8], ARCHS4[9] и GEO Profiles[10]. Как видно из таблицы 3 все сервисы выполняют разные функции, принимая на вход различные аргументы. Сервис GeneQueryDE делает свою уникальную работу – ищет модули дифференциальной экспрессии генов, перекрывающиеся с запросом пользователя, и выдает их, ранжируя в порядке статистической значимости.

Выводы по главе 5

В данной главе были подведены итоги созданного сервиса GeneQueryDE, а именно произведено его сравнение с альтернативной версией по количеству обрабатываемых данных и сравнение с существующими проектами.

Таблица 3 – Сравнение сервисов

	Входные данные	Обрабатываемые виды	Результаты работы
GeneQuery	Набор генов	— Homo Sapience — Mus Musculus — Rattus Norvegicus	— Перекрытия генов — Тепловые карты экспрессии генов
GeneQueryDE	Набор генов	— Homo Sapience — Mus Musculus — Rattus Norvegicus	— Перекрытия генов — Модули дифференциальной экспрессии
GEOracle	Набор экспериментов	Без ограничений	— Числовые значения дифференциальной экспрессии — Тепловые карты зависимости экспериментов от генов
ARCHS4	— Ген — Набор генов в верхней/-нижней регуляции — (доп.параметры: ткани, органы)	— Homo Sapience — Mus Musculus	— Информация о гене — Средняя экспрессия гена
GEO Profiles	— Ген — Эксперимент — Организм — Свободный текст	Без ограничений	— Эксперименты с аннотациями — Информация о генах

ЗАКЛЮЧЕНИЕ

В результате данной работы:

— Построены модули дифференциальной экспрессии генов для видов Homo Sapience, Mus Musculus, Rattus Norvegicus.

— Расширен сервис GeneQuery до версии GeneQuery-de, обеспечивающий интерфейс доступа к модулям генов.

— В версии GeneQuery-de значительно увеличено количество обрабатываемых данных по сравнению с GeneQuery.

— Контейнеризация полученного решения.

— Сервис GeneQuery-de доступен по ссылке: *http://genome.ifmo.ru/genequery-de*.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1 *А.Браун Т.* Геномы. — М.-Ижевск : Институт компьютерных исследований, 2011. — С. 921.

2 *N.Yu. Oparina A.V. Snezhkina A. S.* Differential expression of genes that encode glycolysis enzymes in kidney and lung cancer in humans // Russian Journal of Genetics. — 2013. — Vol. 49. — P. 707–716.

3 *Edgar R Domrachev M L. A.* Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. // Nucleic Acids Res. — 2002. — Vol. 30. — P. 207–210.

4 GeneQuery. — 2016. — URL: `http://genome.ifmo.ru/genequery/searcher/`.

5 Bioconductor. — 2018. — URL: `https://www.bioconductor.org/`.

6 Bonferroni correction. — 2018. — URL: `https://en.wikipedia.org/wiki/Bonferroni_correction`. [Электронный ресурс].

7 Dockerfile reference. — 2018. — URL: `https://docs.docker.com/engine/reference/builder/`.

8 GEOracle. — 2017. — URL: `http://georacle.victorchang.edu.au/`.

9 ARCHS4: Massive Mining of Publicly Available RNA-seq Data from Human and Mouse. — 2018. — URL: `https://amp.pharm.mssm.edu/archs4/index.html`.

10 GEO Profiles. — 2018. — URL: `https://www.ncbi.nlm.nih.gov/geoprofiles/`.