ISSN 2079-3316 ПРОГРАММНЫЕ СИСТЕМЫ: ТЕОРИЯ И ПРИЛОЖЕНИЯ №1(24), 2015, с. 157–174

УДК 004.031.2

А. М. Спицина, Ю. Л. Орлов, Н. Н. Подколодная,
А. В. Свичкарев, А. И. Дергилев, М. Чен, Н. В. Кучин,
И. Г. Черных, Б. М. Глинский

Суперкомпьютерный анализ геномных и транскриптомных данных, полученных с помощью технологий высокопроизводительного секвенирования ДНК

Аннотация. Развитие технологий высокопроизводительного секвенирования ДНК привело к появлению нового класса объемных геномных данных и алгоритмов их обработки и анализа. Суперкомпьютерные вычисления являются необходимым инструментом работы с генетическими данными. Представлены задачи геномики и транскриптомики, анализа экспрессии генов в контексте вычислительной сложности. Дан обзор компьютерных подходов и разработанных авторами программ для решения задач, возникающих при аннотации геномных данных и анализе экспрессии генов.

Ключевые слова и фразы: биоинформатика, секвенирование ДНК, микрочипы, регуляция экспрессии генов, транскрипция, базы данных.

Введение

Стремительное развитие современных молекулярно-биологических и геномных технологий ведет к бурному росту объемов данных высокопроизводительного секвенирования ДНК, что требует развития адекватных компьютерных методов анализа таких данных, опирающихся на суперкомпьютерные технологии. С начала 2000-х

Работа поддержана бюджетным проектом ИЦиГ СО РАН VI.61.1.2, Интеграционным проектом СО РАН и РФФИ (14-04-01906 и 15-54-53091)[2, 6, 7, 8, 9, 9].

© А. М. Спицина[1], Ю. Л. Орлов[2], Н. Н. Подколодная[3], А. В. Свичкарев[4], А. И. Дергилев[5], М. Чен[6], Н. В. Кучин[7], И. Г. Черных[8], Б. М. Глинский[9], 2015
© Институт цитологии и генетики СО РАН[1, 2, 3], 2015
© Новосибирский государственный университет[4, 5], 2015
© Университет Чжецзянь, г. Ханчжоу, Китай[6], 2015
© Институт вычислительной математики и математической геофизики СО РАН[7, 8, 9], 2015
© Программные системы: теория и приложения, 2015

годов, после секвенирования первых полных геномов в молекулярной генетике произошла технологическая революция, связанная с появлением экспрессионных микрочипов высокой плотности и технологий массового параллельного секвенирования ДНК. В связи с этим встает ряд объемных задач анализа геномных данных, включая разработку специализированного программного обеспечения [1]. Полногеномная аннотация кроме определения положения и структуры белок-кодирующих генов, включает описание некодирующих РНК, выделение регуляторных районов генов, исследование однонуклеотидных полиморфизмов, предсказание их вторичной и пространственной структуры белков [2].

Современные методы секвенирования ДНК позволяют не только измерять уровни транскрипции генов (количество мРНК) в клетке, но и решать качественно новые научные проблемы организации генома. Особое место среди методов, основанных на иммунопреципитации хроматина (ChIP) и последующем секвенировании, занимает метод ChIA-PET (Chromatin Interaction Analysis by Paired-End-Tag sequencing), который позволяет исследовать не только отдельные сайты связывания, но пары таких сайтов на районах хромосом, контактирующих в трехмерном пространстве ядра клетки. В последние годы с использованием методов Hi-C, ChIA-PET получены новые знания об особенностях трехмерной архитектуры (укладки) генома человека в интерфазном ядре клетки, влияющих на регуляцию экспрессии генов [3, 4]. С помощью собственных компьютерных программ была обработана информация о хромосомных контактах, опосредованных транскрипционным фактором ER и комплексом РНК-полимеразы II, полученная с помощью метода ChIA-PET [5, 6]. Показано, что геномные области хромосомных контактов, опосредованных комплексом РНК-полимеразы II, обогащены сайтами связывания транскрипционных факторов (полученных по данным ChIP-seq в проекте ENCODE), и участками модификаций гистонов, связанными с активацией экспрессии генов.

Исследование регуляции экспрессии генов в масштабе генома требует развития программных средств интеграции данных, включая данные RNA-seq, ChIP-seq, Hi-C, так же как и микрочиповых данных [4, 7]. В ИЦиГ СО РАН разработан ряд программных средств такой интеграции данных [1, 8–10]. Анализ проводился на вычис-

лительных ресурсах Сибирского Суперкомпьютерного Центра СО РАН.

1. Измерение экспрессии генов

Широкое применение для компьютерного анализа экспрессии генов получили биочипы, или ДНК-микрочипы [11]. Они применяются в самых различных областях современной биологии и медицине, для анализа сложных смесей ДНК в том числе, как небольшого числа проб, так и больших наборов (тысячи проб, вплоть до совокупности всех транскриптов (матричных РНК) в клетке). ДНК-микрочипы используют для анализа изменения экспрессии генов, выявления однонуклеотидных полиморфизмов, генотипирования или повторного секвенирования мутантных геномов.

1.1. Экспрессионные микрочипы

Для решения задач оценки экспрессии генов существует несколько технологических платформ, одна из наиболее распространенных — разработанные компанией Affymetrix микрочипы, использующие технологию синтеза коротких олигонуклеотидных зондов на поверхности микрочипа. Такие данные требуют адекватных статистических и математических методов обработки. Несмотря на широкий спектр методов и компьютерных инструментов, ряд статистических аспектов анализа микрочиповых данных, особенно связанных с интеграцией разрозненных гетерогенных данных, до сих пор не реализован в доступных компьютерных программах. Отметим, в частности, в базы данных BioGPS [12] и Gene Expression Omnibus (GEO) NCBI.

Исходный дизайн олигонуклеотидных проб микрочипа Affymetrix U133, разработанного уже более 10 лет назад, может не соответствовать целевому транскрипту (гену-мишени) и содержать ряд технических проблем, связанных с современной геномной аннотацией. Это ведет к противоречивым результатам, поэтому для работы с данными экспрессии требуется разработка специального программного комплекса, который позволил бы отфильтровать зашумленные сигналы экспрессии на микрочипе и упростить работу с большим объемом данных [13, 14].

Для более полного и точного решения задач, возникающих при анализе экспрессии генов, используются высокопроизводительные

компьютеры. Разработаны различные программные комплексы и надстройки для теоретического, статистического анализа данных микрочипов [**13**], MicroArray DAta Manager (MADAM).

Среди практических приложений отметим исследование экспрессии генов человека. Высокая экспрессия ряда генов может служить маркером для диагностики раковых заболеваний [**11**].

1.2. Микрочипы Affymetrix

Прикладные задачи анализа экспрессии генов состояли в выявлении особенностей генов, активно экспрессирующихся в тканях мозга человека, генов, функционирующих в составе известных генных сетей, и анализе особенностей экспрессии пар транскриптов, ко-локализованных в геноме, в том числе цис-антисенс транскриптов [**15**].

Для выполнения задач реализован инструментарий для обработки данных микрочипов Affymetrix U133 на языке C++, включающий алгоритмы оценок коэффициентов корреляции и фильтрации проб по качеству, т.е. по адекватности измерения уровня экспрессии генов.

Технология синтеза коротких олигонуклеотидных зондов (25 пар нуклеотидов) непосредственно на поверхности микрочипа in situ с использованием литографических масок была разработана компанией Affymetrix для изготовления микрочипов GeneChip. Олигонуклеотидная матрица GeneChip использует наборы синтезированных in situ олигонуклеотидных проб, по 11–20 проб в наборе, каждая размером 25 нуклеотидов, для представления транскриптов генов или их изоформ. Для каждого гена-мишени использованы фрагменты-представители (initial target sequences) длиной 150–450 п.н. для выбора и локализации олигонуклеотидных проб. Сигнал от пробы с совершенным совпадением всех нуклеотидов учитывается после вычитания неспецифического сигнала кросс-гибридизации от пробы с одним центральным несовпадающим нуклеотидом (Affymetrix, 2002).

1.3. Базы данных микрочипов

Стремительное развитие микрочиповых технологий привело к возникновению огромного количества данных, полученных в результате экспериментов по измерению экспрессии генов. Здесь возникают задачи, решением которых занимаются программисты и математики — разработка пакетов для хранения и упорядочивания информации.

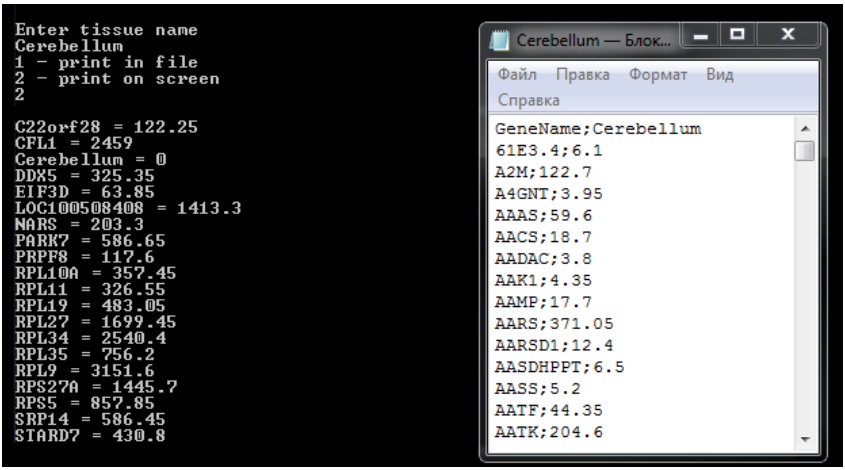


Рис. 1. Опция выдачи значений экспрессии генов из БД по имени ткани: выдача на экран и в файл

Для работы с данными экспрессии генов на микрочипах Affymetrix была выбран портал BioGPS [12]. База данных BioGPS компилирует данные для экспрессии генов человека более чем для 80 видов тканей. Кроме того, данные в BioGPS находятся в свободном доступе. Использовались данные Affymetrix по человеку, но есть и другие типы данных — геномные наборы мыши и крысы, которые можно анализировать с помощью разработанной программы.

Набор микрочипа Affymetrix U133 — это комплект, состоящий из двух массивов. Он содержит порядка 45 000 наборов проб, представляющих более чем 39 000 транскриптов, полученных из примерно 33 000 аннотированных в геноме генов человека.

Сложность работы с данными в таком виде заключается в их объеме — 45 000 строк и 80 столбцов, в отсутствии единого формата (часть таблицы — текстовые имена, а часть — числовые данные), а также в том, что каждая строка — это отдельный элемент с набором параметров (проба гена, его идентификаторы и его экспрессия). Некоторые гены имеют по нескольку соответствующих им строк — пробы-дубли, которые осложняют анализ. На рис. 1 показан пример выдачи значений экспрессии генов из БД по имени ткани, выдача на экран и в файл (для ткани «cerebellum» — мозжечок).

Ранее были опубликованы работы по оценке качества проб [**13**], позволяющие отфильтровать пробы по качеству и использовать для оценки экспрессии гена только те пробы микрочипа, которые для данного гена однозначно соответствуют транскрипту РНК. Разработанная компьютерная программа позволяет использовать такой фильтр при выборе из нескольких проб.

1.4. Пример разработанной компьютерной программы

Некорректность и большой объем экспериментальных данных требуют значительного количества времени для анализа и обработки, при этом велика вероятность получения ошибок, связанных с человеческим фактором. Так, например, построенная матрица корреляции в уже отфильтрованной базе данных имеет размер порядка 20000×20000, и для ее подсчета и выдачи в файл потребовалось около суток времени вычислений на персональном компьютере.

Программный комплекс должен отвечать следующим требованиям:

- хранение базы данных в памяти, чтобы с ними можно было работать последовательно, не считывая их каждый раз из файла. Это сэкономит время доступа к базе данных, а также позволит работать с измененными данными (например, после объединения нескольких баз данных);
- простой и понятный интерфейс.

В различных базах данных одному гену может сопоставляться несколько параметров, и для статистического анализа возникает необходимость сопоставлять данные из разных источников. Вручную сопоставить несколько таблиц размером 20 000–40 000 строк-проб, с данными, которые могут быть не полными или не всегда совпадают по идентификаторам — очень трудоемкая задача. Чтобы ускорить процесс обработки и повысить точность результатов, необходима разработка универсального программного комплекса для работы с данными в общем виде, который будет работать с каждой пробой-строкой как с отдельным элементом (рис. 2) и выполнять такие функции, как сопоставление проб (нахождение максимального или среднего значения), получение информации по задаваемой ткани или имени гена, объединение данных из различных баз, удаление проб-дублей, и

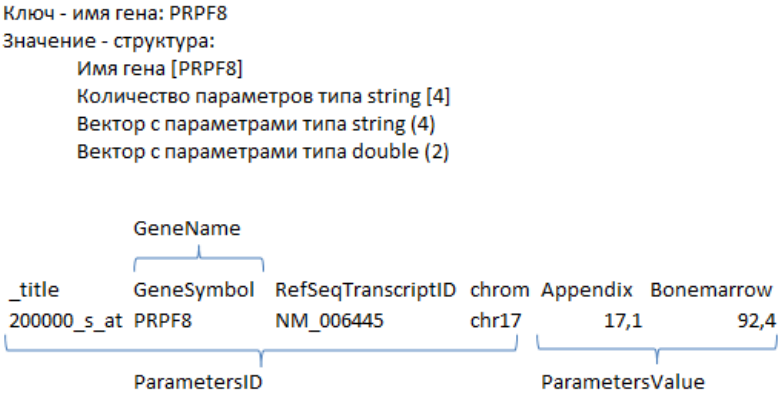


Рис. 2. Пример описания структуры из разработанной программы, в которой хранится проба микрочипа, идентификатор гена и значения уровней экспрессии гена в тканях организма

построение матрицы корреляций по списку генов для статистического анализа.

Кроме того, программный комплекс должен быть совместим с разработанными ранее в ИЦиГ СО РАН блоками и расчетными модулями (такими как JACOBI 4) [16].

2. Корреляции экспрессии генов

С помощью компьютерного анализа данных микрочипов можно выделять группы генов, которые дифференциально экспрессируются в исследуемых тканях организма, а также анализировать связи генов друг с другом используя информацию из генных сетей, представленную в базах данных и научной литературе.

Для этого была разработана программа на языке C++ (порядка 1400 строк), которая позволяет работать с базой данных BioGPS, а также с другими базами данных в текстовом формате, содержащая такие опции, как:

- нахождение пробы с наибольшей/средней экспрессией по имени заданного гена,
- выбор генов с наибольшей экспрессией в заданной ткани,

- вставка информации о генах из другой базы данных,
- фильтрация проб-дублей в базе данных,
- построение матрицы корреляций (линейной или ранговой) по группе генов для одной выборки или сравнение нескольких выборок (заданные гены / случайная выборка), «усредненная» выборка (статистика по заданному числу случайных выборок), а также статистика по ней — гистограмма распределений коэффициентов корреляции и график расположения генов на хромосомах.

2.1. Инструменты программы

Важная часть программы — подсчет матрицы корреляций. Для этого исходные данные фильтруются — удаляются пробы-дубли, чтобы каждому гену соответствовала одна строка. Матрица может считаться по заданному списку генов или по случайной выборке, можно сравнивать две матрицы. Есть возможность посчитать линейные коэффициенты корреляции (коэффициенты Пирсона) или ранговые (коэффициенты Спирмена).

Вычисления происходят следующим образом: из ассоциативного контейнера, хранящего базу данных, случайным образом выбирается заданное число N генов-строк, затем каждый элемент матрицы будущей C[I][J](N×N) считается следующим образом: попарно считается корреляция между наборами значений экспрессии двух генов I и J на выборке тканей. Здесь N меняется от 1 до 20000, I и J — от 1 до 80. На рисунке показана выдача программы анализа матриц корреляции экспрессии для двух выборок генов человека (по 5 генов в каждой), и представление коэффициентов корреляции в форме гистограммы (рис. 3).

Блоками последовательно обозначены рассчитанные матрицы для первой и второй выборок (first sample, second sample), число положительных/отрицательных коэффициентов корреляции (number of coefficients) и гистограммы распределения значений коэффициентов (bar graph).

Для подсчета ранговой корреляции необходимо каждому элементу векторов I и J присвоить ранги, то есть те номера, которые бы имели элементы при упорядочивании по возрастанию. Причем, если несколько элементов совпадают, то номера «усредняются» — их номера складываются и делятся на количество совпавших. Для хранения

```
       First sample:
    ;EPYC; KEL; NDRG1; RNF141; TGDS;
EPYC; 1; 0.0391508; 0.0880769; -0.0298825; -0.0674657;
KEL; 0.0391508; 1; -0.0409506; 0.0117414; -0.0329537;
NDRG1; 0.0880769; -0.0409506; 1; -0.0879477; 0.087617;
RNF141; -0.0298825; 0.0117414; -0.0879477; 1; -0.0459731;
TGDS; -0.0674657; -0.0329537; 0.087617; -0.0459731; 1;

      Second sample:
    ;C6orf25; HBEGF; HDC; LRCH3; MYT1L;
C6orf25; 1; 0.799116; 0.0815886; 0.949629; 0.0443402;
HBEGF; 0.799116; 1; 0.0784912; 0.783506; -0.0254945;
HDC; 0.0815886; 0.0784912; 1; 0.0693734; 0.188815;
LRCH3; 0.949629; 0.783506; 0.0693734; 1; -0.0546877;
MYT1L; 0.0443402; -0.0254945; 0.188815; -0.0546877; 1;

Number of coefficients: 10 / 10
Positive: 4(40%) / 8(80%)
Negative: 6(60%) / 2(20%)
Greater than 0.8: 0(0%) / 1(10%)


Bar graph:
-1;-0.9;-0.8;-0.7;-0.6;-0.5;-0.4;-0.3;-0.2;-0.1;0;0.1;0.2;0.3;0.4;0.5;0.6;0.7;0.8;0.9;1;
0;0;0;0;0;0;0;0;6;0;4;0;0;0;0;0;0;0;0;
0;0;0;0;0;0;0;0;2;0;4;1;0;0;0;0;2;0;1;

Location on chromosomes:
c1;c2;c3;c4;c5;c6;c7;c8;c9;c10;c11;c12;c13;c14;c15;c16;c17;c18;c19;c20;c21;c22;cX;cY
0;0;0;0;0;1;1;0;0;1;1;1;0;0;0;0;0;0;0;0;0;0;0
0;1;1;0;1;1;0;0;0;0;0;0;0;1;0;0;0;0;0;0;0;0;0
```

Рис. 3. Выдача программы анализа матриц корреляции экспрессии для двух выборок генов

рангов заводятся два массива RANK1 и RANK2, в каждом из которых i-ому элементу RANK1[i], RANK2[i] соответствует ранг i-ого элемента I[i], J[i].

3. Исследование генных сетей с помощью корреляций экспрессии генов

3.1. Расчет корреляций для заданных наборов генов

Целью исследования являлось изучение биологических функций выделенных генов в структуре генных сетей, их взаимосвязей, практическим результатом — программа для анализа экспрессионных данных, состоящая из нескольких модулей, работающих с текстовой базой данных (выбор данных по запросу, статистический анализ — корреляция).

В нашей работе реализована программа анализа корреляций, которая упростит выявление структурных особенностей генов с высокой экспрессией, выполняет анализ экспрессии генов человека, анализ различных генных сетей (комплексов взаимодействующих макромолекул в клетке), исследование качества измерения сигнала на микрочипах, анализ тканеспецифичности экспрессии генов.

С помощью разработанной программы среди проб микрочипа Affymetrix U133, представленных в БД BioGPS, были выделены пробы с высокой экспрессией так, чтобы было соответствие «одна проба — один ген». Создан программный инструмент, позволяющий для данной БД строить таблицу корреляций экспрессии пар генов (заданных или выбранных случайно). Из списка тканей были выделены ткани мозга и подготовлены выборки генов, экспрессия которых повышена в структурах мозга. На основе подготовленных выборок выявлены структурные особенности генов с высокой экспрессией (число экзонов, длина транскрипта, связь с альтернативным сплайсингом) [**15**].

С использованием данной программы было проведено исследование генных сетей. В частности, были проанализированы генные сети циркадного ритма (рис. 4) и регуляции холестерина (исследованием генных сетей занимается Отдел системной биологии ИЦиГ СО РАН).

Интересно отметить повышенную фракцию негативных коэффициентов корреляции между уровнями экспрессии генов входящих в генную сеть циркадного ритма, что свидетельствует об отрицательных обратных связях во взаимодействии генов (через их белковые продукты). Также были выделены пробы с высокой экспрессией, подготовлены выборки генов, экспрессия которых повышена в структурах мозга, построены генные сети данных генов.

Гены, имеющие высокий уровень экспрессии в широком круге органов, имеют высокий уровень экспрессии и в структурах головного мозга [**2**]. Был проведен следующий сравнительный анализ. На основе данных об экспрессии генов в различных органах по базе данных BioGPS, был проведен следующий сравнительный анализ. Для каждого гена определялись наиболее высокие значения, лежащие вне доверительного интервала 99%. Если они обнаруживались хотя бы для одной ткани головного мозга, то такие гены группировались. Всего в эту группу вошло 55 генов, имеющих повышенную экспрессию во всех изучаемых тканях.

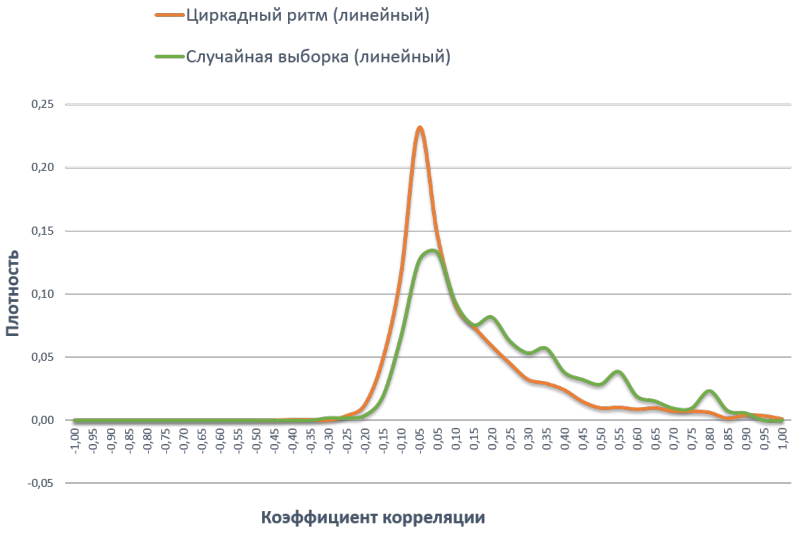


Рис. 4. Гистограммы сравнения генов ГС циркадного ритма и случайной выборки такого же размера

3.2. Развитие программ анализа данных экспрессии генов

Для изучения генома человека с помощью данных экспрессии был создан набор компьютерных программ для работы с существующими базами данных. Программы, реализованные на языке C++, предназначены для статистической обработки данных микрочипов Affymetrix U133 и включают в себя алгоритмы для оценки коэффициентов корреляции и фильтрации проб генов, для выявления особенностей экспрессии генов человека.

Для анализа взаимосвязей между экспрессией генов из генных сетей в различных тканях использовалась база данных GeneNet, содержащая аннотированные генные сети. Полученные с помощью данного инструмента результаты могут быть использованы для дальнейшего исследования генных сетей и метаболических путей.

Данный инструмент будет объединен с программно-алгоритмическим комплексом для многомерного анализа микрочиповых данных JACOBI 4, предназначенным для потоковой обработки похожих данных

одним и тем же алгоритмом [**16**]. Пакет JACOBI 4 представляет собой набор программ для многомерного анализа с открытым кодом, который одинаково удобен для использования пользователями с любым опытом работы с ПК. Проект JACOBI 4 развивается для поддержки новой технологии поиска генов-кандидатов в генные сети, разработанной в ИЦиГ СО РАН, и для расширения его функциональности требуется интеграция инструмента для обработки данных Affymetrix.

4. Задачи компьютерной аннотации геномных данных

Важнейшим объектом геномики являются молекулярно-генетические системы, координирующие функции генов, РНК, белков, которые можно исследовать на уровне транскрипции через измерение экспрессии генов, как на микрочипах, так и с помощью транскриптомного секвенирования. Сложность анализа только увеличивается при рассмотрении взаимодействий между генами (в форме корреляций), исследовании сетевых взаимодействий в метаболических путях. Несмотря на распространение компьютерных программ биоинформатики, остается ряд направлений развития программного обеспечения, требующих более детальной алгоритмической разработки и реализации в форме специализированного программного обеспечения на различных вычислительных платформах. Можно выделить следующие направления анализа геномных данных, связанных с секвенированием и экспрессией генов:

(1) Разработка конвейерного подхода для процессинга, картирования на референсный геном последовательностей, полученных в ходе экспериментов секвенирования (включая данные RNA-seq).
(2) Функциональная аннотация генома человека и модельных организмов на основе интеграции данных о положении регуляторных районов транскрипции генов (по данным ChIP-seq и родственных технологий).
(3) Разработка программ для анализа структур РНК, некодирующих РНК и миРНК, разметки их функциональных сайтов.
(4) Анализ трансляции генов, регуляции экспрессии на уровне трансляции, определения свойств белковых фрагментов, кодируемых в нуклеотидных последовательностях
(5) Сравнение функциональных свойств вновь секвенированных генов различных организмов (задачи сравнительной геномики).

Решение этих задач необходимо для обеспечения технической поддержки геномных исследований. Ранее технические средства этого назначения реализованы в разработанном программном комплексе ICGenomics, представленном на ССКЦ СО РАН (ЦКП «Биоинформатика» СО РАН. Особое внимание было уделено оригинальным методам, не повторяющим стандартные алгоритмы, таких, как предсказание сайтов связывания транскрипционных факторов (ССТФ) по нуклеотидной последовательности (с помощью весовых матриц).

Следует отметить, что за короткий период в последние 2–3 года на смену микрочипам приходят все более совершенные технологии полного секвенирования транскриптом (RNA-seq), имеющие ряд преимуществ, в частности, по способности определения новых вариантов транскриптов, по динамической шкале измерения уровня транскрипции. Таким образом, задачи компьютерного анализа геномных последовательностей, объединяемых общими типами данных, требуют дальнейшего развития. Происходит и объединение ресурсов с иностранными партнерами, в частности в области разработки геномных баз данных.

Заключение

Разработан программный комплекс анализа экспрессии генов, использующий ряд уникальных модулей. Программа позволяет выполнять ряд функций обработки и анализа геномных последовательностей: Исследовано распределение уровней экспрессии генов по микрочиповым данным БД BioGPS. Выявлены особенности корреляций между генами в составе генных сетей.

Благодарности. Авторы благодарны Н. Л. Подколодному, Е. В. Кулаковой, Н. С. Сафроновой, Х. Бай, коллегам из Национального Университета Внутренней Монголии КНР, а также ССКЦ СО РАН за поддержку работы.

Список литературы

[1] Ю. Л. Орлов, А. О. Брагин, И. В. Медведева и др. «ICGenomics: программный комплекс анализа символьных последовательностей геномики», *Вавиловский журнал генетики и селекции*, **16**:4/1 (2012), с. 732–741, URL http://vavilov.elpub.ru/index.php/jour/article/view/70 ↑ 158.

[2] J. C. Kwasnieski, C. Fiore, H. G. Chaudhari, B. A. Cohen. «High-throughput functional testing of ENCODE segmentation predictions», *Genome Res.*, **24**:10 (2014), с. 1595–1602, URL http://www.ncbi.nlm.nih.gov/pubmed/25035418 ↑ 158, 166.

[3] F. Ay, T. L. Bailey, W. S. Noble. «Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts», *Genome Res.*, **24**:6 (2014), с. 999–1011, URL http://www.ncbi.nlm.nih.gov/pubmed/24501021 ↑ 158.

[4] Ю. Л. Орлов. «Компьютерное исследование регуляции транскрипции генов эукариот с помощью данных экспериментов секвенирования и иммунопреципитации хроматина», *Вавиловский журнал генетики и селекции*, **18**:1 (2014), с. 193–206 ↑ 158.

[5] M. J. Fullwood, M. H. Liu, Y. F. Pan et al. «An oestrogen-receptoralpha-bound human chromatin interactome», *Nature*, **462**:7269 (2009), с. 58–64, URL http://www.ncbi.nlm.nih.gov/pubmed/19890323 ↑ 158.

[6] G. Li, X. Ruan, R. K. Auerbach et al. «Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation», *Cell*, **148**:1–2 (2012), с. 84–98, URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3339270/ ↑ 158.

[7] Y. Orlov, H. Xu, D. Afonnikov et al. «Computer and Statistical Analysis of Transcription Factor Binding and Chromatin Modifications by ChIP-seq data in Embryonic Stem Cell», *J. Integr. Bioinform.*, **9**:2 (2012), с. 211, URL http://www.ncbi.nlm.nih.gov/pubmed/22987856 ↑ 158.

[8] О. С. Кожевникова, М. К. Мартыщенко, М. К. Генаев и др. «RatDNA: база данных микрочиповых исследований на крысах для генов, ассоциированных с заболеваниями старения», *Вавиловский журнал генетики и селекции*, **16**:4/1 (2012), с. 756–765, URL http://vavilov.elpub.ru/index.php/jour/article/view/72 ↑ 158.

[9] И. В. Медведева, О. В. Вишневский, Н. С. Сафронова и др. «Компьютерный анализ данных экспрессии генов в клетках мозга, полученных с помощью микрочипов и высокопроизводительного секвенирования», *Вавиловский журнал генетики и селекции*, **17**:4/1 (2013), с. 629–638, URL http://vavilov.elpub.ru/index.php/jour/article/view/187 ↑ 158.

[10] А. М. Спицина, «Компьютерное исследование экспрессии генов человека с использованием базы данных BioGPS микрочипов Affymetrix U133», *Студент и научно-технический прогресс*, Материалы 52-й международной научной студенческой конференции, НГУ, Новосибирск, 2014, URL http://issc.nsu.ru/wp-content/uploads/2014/11/07Biology.pdf ↑ 158.

[11] A. Perez-Diez, A. Morgun, N. Shulzhenko. “Microarrays for cancer diagnosis and classification”, *Adv. Exp. Med. Biol.*, **593** (2013), pp. 74–85, URL http://www.ncbi.nlm.nih.gov/books/NBK6624/ ↑ 159, 160.

[12] C. Wu, C. Orozco, J. Boyer et al. “BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources”, *Genome Biol.*, **10**:11 (2009), pp. R130, URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3091323/ ↑ 159, 161.

[13] Y. L. Orlov, J. Zhou, L. Lipovich et al. “Quality assessment of the Affymetrix U133A&B probesets by target sequence mapping and expression data analysis”, *Silico Biol.*, **7**:3 (2007), pp. 241–60, URL http://www.ncbi.nlm.nih.gov/pubmed/18415975 ↑ 159, 160, 162.

[14] Ю. Л. Орлов, В. М. Ефимов, Н. Г. Орлова. «Статистические оценки экспрессии мобильных элементов в геноме человека на основе клинических данных экспрессионных микрочипов», *Вавиловский журнал генетики и селекции*, **15**:2 (2011), с. 327–339, URL http://www.bionet.nsc.ru/vogis/pict_pdf/2011/15_2/12.pdf ↑ 159.

[15] И. В. Медведева, О. В. Вишневский, Н. С. Сафронова и др, «Геномная организация и контекстные характеристики генов с повышенной экспрессией в клетках мозга», *Нейроинформатика-2014*, Сборник научных трудов. Часть 2, XVI Всероссийская научно-техническая конференция, НИЯУ МИФИ, М., 2014, с. 32–42 ↑ 160, 166.

[16] Д. А. Полунин, И. А. Штайгер, В. М. Ефимов. «Разработка программного комплекса JACOBI 4 для многомерного анализа микрочиповых данных», *Вестник НГУ. Серия: Информационные технологии*, **12**:2 (2014), с. 90–98, URL http://www.nsu.ru/xmlui/bitstream/handle/nsu/4125/2014_V12 ↑ 163, 168.

Рекомендовал к публикации *Рекомендована к публикации Программным комитетом НСКФ-2014*

Об авторах:

Анастасия Михайловна Спицина

Магистрант Новосибирского государственного университета. Область научных интересов: биоинформатика, суперкомпьютерные вычисления.

e-mail: anastasia.spitsina@gmail.com

Юрий Львович Орлов

Окончил НГУ в 1991 г., д.б.н., с.н.с., зав. лабораторией компьютерной геномики ФЕН НГУ, зав.лаб. нейроинформатики поведения ИЦиГ СО РАН. Область научных интересов: биоинформатика, компьютерная геномика.

e-mail: orlov@bionet.nsc.ru

Наталья Николаевна Подколодная

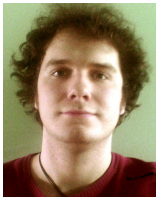
Окончила Новосибирский государственный университет, научный сотрудник ИЦиГ СО РАН. Область научных интересов: генные сети, суперкомпьютерные вычисления.

e-mail: nata@bionet.nsc.ru

Анатолий Владленович Свичкарев

Студент НГУ. Область научных интересов: биоинформатика, геномика, суперкомпьютерные вычисления.

e-mail: tolik0393@mail.ru



Артур Игоревич Дергилев

Студент НГУ. Область научных интересов: биоинформатика, суперкомпьютерные вычисления.

e-mail: arturd1993@yandex.ru

Минг Чен

Профессор, зав. лабораторией биоинформатики Колледжа Естественных Наук, Университет Чжецзянь, г.Ханчжоу, Китай. Область научных интересов: биоинформатика, моделирование генных сетей, интеграция данных.

e-mail: mchen@zju.edu.cn

Николай Владимирович Кучин

Окончил НГУ в 1971 г., главный специалист по системному программному обеспечению ИВМиМГ СО РАН. Область интересов: высокопроизводительные вычислительные системы, системное программное обеспечение кластеров

e-mail: kuchin@sscc.ru

Игорь Геннадьевич Черных

Окончил НГУ в 2002 г., кандидат физико-математических наук. Область научных интересов: суперкомпьютерные вычисления, химическая кинетика.

e-mail: chernykh@parbz.sscc.ru

Борис Михайлович Глинский

Окончил НГУ в 1967 г., профессор, доктор технических наук. Область научных интересов: вычислительные системы, моделирование сейсмических полей, имитационное моделирование.

e-mail: gbm@sscc.ru

Пример ссылки на эту публикацию:

А. М. Спицина, Ю. Л. Орлов и др.. «Суперкомпьютерный анализ геномных и транскриптомных данных, полученных с помощью технологий высокопроизводительного секвенирования ДНК», *Программные системы: теория и приложения*, 2015, **6**:1(24), с. 157–174.

URL http://psta.psiras.ru/read/psta2015_1_157-174.pdf

Anastasiya Spitsina, Yurij Orlov, Natalya Podkolodnaya, Anatolij Svichkarev, Artur Dergilev, Ming Chen, Nikolaj Kuchin, Igor Chernykh, Boris Glinskij. *Supercomputer analysis of genomics and transcriptomics data revealed by high-throughput DNA sequencing.*

Abstract. Development of high-throughput DNA sequencing technologies lead to new classes of bulk genomic data and consequent development of specialized algorithms and software. Supercomputing is necessary tool to deal with modern genetics data. We present technical problems related to gene expression analysis, genomics and transcriptomics data, as well as sequencing technologies related to gene expression. Approaches for automatic genome data annotation are discussed. (*In Russian*).

Key Words and Phrases: Bioinformatics, DNA sequencing, Microarrays, Gene expression regulation, Transcription, Databases.

Sample citation of this publication

A. M. Spitsina, Yu. L. Orlov et al.. "Supercomputer analysis of genomics and transcriptomics data revealed by high-throughput DNA sequencing", *Program systems: theory and applications*, 2015, **6**:1(24), pp. 157–174. (*In Russian.*)

URL http://psta.psiras.ru/read/psta2015_1_157-174.pdf

© A. M. Spitsina[1], Y. L. Orlov[2], N. N. Podkolodnaya[3], A. V. Svichkarev[4], A. I. Dergilev[5], M. Chen[6], N. V. Kuchin[7], I. G. Chernykh[8], B. M. Glinskij[9], 2015
© Institute of Cytology and Genetics SB RAS[1, 2, 3], 2015
© Novosibirsk State University[4, 5], 2015
© Zhejiang University[6], 2015
© Institute of Computational Mathematics and Mathematical Geophysics SB RAS[7, 8, 9], 2015
© Program systems: Theory and Applications, 2015