ISSN 2079-3316 ПРОГРАММНЫЕ СИСТЕМЫ: ТЕОРИЯ И ПРИЛОЖЕНИЯ №2(25), 2015, с. 129–148

УДК 004.031.2

Е. В. Кулакова, А. М. Спицина, Н. Г. Орлова, А. И. Дергилев,
А. В. Свичкарев, Н. С. Сафронова, И. Г. Черных, Ю. Л. Орлов

Программы анализа геномных данных секвенирования, полученных на основе технологий ChIP-seq, ChIA-PET и Hi-C

Аннотация. Возрастающие объемы геномных данных о положении сайтов связывания транскрипционных факторов, хромосомных контактах, аннотации геномных характеристик, полученных с помощью современных технологий секвенирования, требуют разработки нового программного обеспечения для их анализа, оптимизации существующих алгоритмов обработки. Суперкомпьютерные вычисления позволяют решать задачи исследования регуляции транскрипции генов на качественно новом уровне. Рассмотрены задачи анализа геномных данных секвенирования, полученных на основе технологий ChIP-seq, ChIA-PET и Hi-C. Представлены компьютерные подходы и разработанные авторами программы для решения предложенных задач геномики, приведена дискуссия о дальнейших направлениях развития.

Ключевые слова и фразы: биоинформатика, секвенирование ДНК, иммунопреципитация, хромосомные контакты, регуляция экспрессии генов, базы данных.

Введение

Суперкомпьютерные вычисления позволяют решать задачи исследования регуляции транскрипции генов на качественно новом уровне. Продолжающееся в последние годы стремительное развитие геномных технологий высокопроизводительного секвенирования ДНК, также называемое «секвенирование следующего поколения» (NGS, или Next Generation Sequencing), ведет к росту объемов доступных геномных данных [1]. Развитие экспериментальных молекулярно-биологических технологий ChIP-seq, связанных с иммунопреципитацией хроматина и

Работа поддержана бюджетным проектом ИЦиГ СО РАН VI.61.1.2. Исследование данных ChIA-PET поддержано грантом РФФИ 14-04-01707.

последующим секвенированием фрагментов ДНК (Chromatin Immuno-Precipitation-Sequencing), позволяет ставить новые масштабные задачи: исследовать сайты связывания транскрипционных факторов (ССТФ) в масштабе генома, определять гены–мишени воздействия этих факторов, сравнивать уровни экспрессии этих генов (активность транскрипции мРНК). Возникают качественно новые постановки задач биологического исследования, включая разработку специализированного программного обеспечения [1, 2]. Рассмотрим применение современных математических и компьютерных методов анализа регуляции транскрипции эукариот с использованием полногеномных данных ChIP-экспериментов, связанных с иммунопреципитацией хроматина и последующим секвенированием (ChIP-PET, ChIP-seq, ChIA-PET) [3]. Такие технологии применяются для картирования сайтов связывания транскрипционных факторов, анализа особенностей организации регуляторных районов генов и структуры хроматина в масштабе генома, выяснения деталей молекулярных механизмов регуляции транскрипции генов.

Исследование данных о сайтах связывания транскрипционных факторов и регуляторных районов генов в масштабе генома требует развития программных средств интеграции данных [3, 4].

Цель работы — создание набора программных утилит для обработки полногеномных данных экспериментов данной серии по выявлению и анализу сайтов связывания транскрипционных факторов с использованием иммунопреципитации хроматина.

Разработана серия программ анализа нуклеотидных последовательностей сайтов (выявление нуклеотидных мотивов, профили сложности текста), анализа расположения сайтов и выявления взаимодействующих генов, а также списков полученных генов в контексте генных сетей [5–7]. Анализ проводился на вычислительных ресурсах Сибирского Суперкомпьютерного Центра СО РАН (http://www2.sscc.ru/).

1. Данные ChIP-seq и Hi-C

Коротко рассмотрим экспериментальный метод ChIP-seq [1], позволяющий изучать связывание транскрипционных факторов (ТФ) с ДНК — метод иммунопреципитации хроматина (Chromatin ImmunoPrecipitation — ChIP). Метод ChIP-seq включает стадию иммунопреципитации (выделения белков, связанных с ДНК с помощью антител) и последующее секвенирование ДНК (посимвольное определение последовательностей ДНК), связанных с этими белками. ДНК,

выделенная путем иммунопреципитации со специфическими антителами, связывающими исследуемый белок (фактор транскрипции), отмывается от белковой фракции. Оставшиеся фрагменты ДНК имеют размер в несколько сот пар оснований (обычно 150-300 п.о.). Секвенирование (определение нуклеотидной последовательности) проводится на приборах массового параллельного секвенирования, производящих от мегабаз до 1 гигабазы символьных последовательностей на один эксперимент.

Распространены технологии секвенирования компаний Roche 454, Illumina и SOLiD, отличающиеся по физическим принципам работы, по форматам данных, а также (условно) по длине определяемой последовательности и максимальному размеру длин прочтений ДНК (20-300 нуклеотидов) [3]. Отметим обратную пропорциональность между длиной последовательностей и вычислительной сложностью обработки данных секвенирования — чем короче последовательности прочтений ДНК, тем сложнее и больше времени занимает картирование — определение однозначного соответствия прочтений ДНК референсному геному (протяженной последовательности, в которой надо найти короткую последовательность прочтения ДНК — от 18 до 100 нуклеотидов), и тем выше вычислительная сложность задачи. Заметим, что для коротких последовательностей (порядка 20 нуклеотидов) может не быть однозначного соответствия в большом геноме (3 млрд. нуклеотидов).

Обработка найденного набора сайтов связывания исследуемого транскрипционного фактора, найденных в геноме при помощи ChIP-seq, требует разработки компьютерных программ сравнения положения таких сайтов и расположения генов в геноме, определения потенциальных генов мишеней действия этого транскрипционного фактора. Технология ChIA-PET (Chromatin Interaction Analysis by Paired-End-Tag sequencing) позволяет исследовать не только отдельные сайты связывания, но пары таких сайтов на районах хромосом, контактирующих в трехмерном пространстве ядра клетки [3]. Рассмотрим соотношение ChIP технологий (рис. 1) по возможностям и объемам данных.

Еще более сложной и объемной становится задача описания регуляторных районов генов и определения генов-мишеней при анализе данных трехмерных контактов по экспериментальным методам Hi-C [8] и ChIA-PET [9, 10]. Данными являются уже не линейные координаты последовательностей, а двумерные — наборы пар контактов

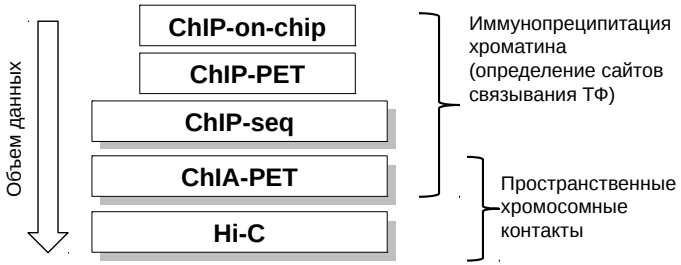

Рис. 1. Соотношения ChIP-технологий по объемам данных и решаемым задачам

и матрицы контактов. Метод Hi-C (аббревиатура от High dimension Chromosome), позволяет определить пространственную структуру хроматина в масштабе всего генома. Метод основан на технологии 3C (Chromosome Conformation Capture) и технологиях массового параллельного секвенирования [11]. Выходными данными Hi-C, также как и ChIA-PET, являются парные последовательности прочтений ДНК (от 20–100 млн. парных прочтений), которые после расположения на геноме позволяют оценить вероятность взаимодействия двух районов генома в большой популяции клеток.

Методы ChIP-on-chip и ChIP-PET аналогичны ChIP-seq и являются скорее ограниченными вариантами ChIP-seq, направленными только на изучение ограниченного числа районов в геноме (промоторов генов), либо использовавшими уже устаревшие методы секвенирования. Изучение хромосомных контактов в ядре клетки важно для анализа экспрессии (функционирования) генов, их регуляции (управления работой) посредством транскрипционных факторов, зачастую связывающихся с ДНК удаленно от генов. Анализ новых трехмерных данных, полученных с помощью полногеномных методов Hi-C и ChIA-PET, позволяет по-новому рассмотреть задачу анализа генных сетей.

Для нашего анализа первичные данные секвенирования поступают в формате bed-файлов. Распространен FASTA формат для картирования на геном (тысячи и миллионы последовательностей). Объем файлов типичного эксперимента ChIP-seq составляет от 100Мб до 1Гб, что задает соответствующие требования к программному обеспечению. Объем данных ChIA-PET — в несколько раз больше.

1.1. Базы данных геномного секвенирования и ресурсы

Среди существующих решений задач, связанных с обработкой полногеномных данных, есть интернет-ресурсы, такие как Galaxy, ChIA-PET Tool [12] и совсем недавно появившийся 3DGD [13]. 3DGD (Three dimensional genome database) — это база данных Hi-C, обеспечивающая доступ и визуализацию трехмерной структуры хроматина и интегрируемая с другими данными масштаба генома, такими как ДНК-связывающий белок. Данные по выделенным доменам Hi-C в геноме мыши доступны на сайте лаборатории проф. B. Ren.

База данных CTCFBSDB 2.0 содержит информацию о расположении хромосомных доменов, ограниченных инсулятором CTCF (CCCTC-binding factor — транскрипционный фактор, определяющий границы транскрипции на хромосоме).

Использовались базы данных BioGPS [14] и Gene Expression Omnibus (GEO) NCBI. BioGPS — это расширяемый портал, содержащий данные в свободном доступе. База данных BioGPS включает данные экспрессии генов человека более чем для 80 видов тканей и типов клеток. Центр биомедицинской информации NCBI разрабатывает информационные технологии для анализа молекулярно-генетических процессов, связанных с заболеваниями человека, и содержит крупнейший репозитарий данных геномного секвенирования. GEO (Gene Expression Omnibus) — это открытое хранилище данных функциональной геномики. GEO содержит инструменты, помогающие пользователям создавать запросы и загружать экспериментальные данные.

Сервис Galaxy — открытая веб-платформа для доступа и воспроизведения вычислений биомедицинских исследований предоставляет инструменты анализа геномных данных, возможность конвертации форматов данных секвенирования. ChIA-PET Tool — пакет программного обеспечения для автоматической обработки данных о последовательностях сайтов, найденных с помощью ChIA-PET, дает отображение результатов на графическом геномном браузере [12].

Проект JACOBI [15] развивается для поддержки технологии поиска генов-кандидатов в генные сети, разработанной в ИЦиГ СО РАН. Среди практических приложений анализа ChIP-данных и данных секвенирования отметим исследование регуляции экспрессии генов человека, анализ полиморфизмов [16]. Определение генов мишеней (участков связывания ТФ в регуляторных районах, контролирующих транскрипцию этих генов) может служить маркером для диагностики

заболеваний [3].

Если рассматривать распределение прочтений ДНК (ридов) из эксперимента (профиль ChIP-seq), то можно увидеть, что оно неравномерно. В отдельных областях ридов будет значительно больше, чем в других. Эти области принято называть пиками. Как правило, они содержат внутри себя сайты связывания ТФ, поскольку были связаны с белком в начале эксперимента. Можно сформировать bed-файл, который представляет собой запись позиций в геноме: хромосомы и границы для каждого пика. В нашей работе за исходные данные брались такие файлы с интервалами, полученные ранее в работах [1] и представленные в ресурсе GEO NCBI. Использовались геномы мыши (версии mm8, mm9, mm10) и человека (hg19).

В данной работе использовались следующие данные:

- Референсный геном из геномного браузера UCSC Genome Browser. Genome Browser позволяет масштабировать и визуально «прокручивать» хромосомы, выделяя нужные районы. Table Browser обеспечивает доступ к основной базе данных. Genome Graphs отображает полногеномные наборы данных.
- Данные о сайтах связывания транскрипционного фактора белка CTCF взяты из CTCFBSDB 2.0 и ресурса Jaspar.
- Список генов, относящихся к контролю агрессивного поведения у лабораторных животных (мыши и крысы), полученные в ИЦиГ СО РАН в рамках работы по проекту РНФ «Экспериментальные генетические модели агрессивного и толерантного поведения: исследование молекулярно-генетических механизмов с использованием технологий секвенирования следующего поколения (RNA-Seq)» [17].

1.2. Средства разработки

Для выполнения задач, связанных с обработкой данных экспрессии, реализован инструментарий для обработки данных на языке C++.

Также был реализован, протестирован и апробирован на экспериментальных данных набор утилит для анализа данных секвенирования. Утилиты реализованы на языке Python версии 2.7, использовались открытые библиотеки Biopython, MOODS, bbcflib. Утилиты предназначены для локализации мотива среди нуклеотидных последовательностей, уточнения позиционных матриц и их визуализации. Набор утилит может быть использован в качестве стадии конвейера.

Задачи, относящиеся к анализу расположения генов, сайтов связывания и доменов были реализованы на языке Java. В качестве среды разработки была выбрана среда программирования NetBeans IDE. JDK последней версии 1.8. Компонент Swing GUI Builder в NetBeans IDE упрощает процесс разработки графического интерфейса и позволяет использовать визуальные инструменты и предварительно установленные компоненты Swing и AWT для создания графического интерфейса приложений Java.

Для автоматической загрузки данных референсного генома из геномного браузера UCSC Genome Browser была использована открытая библиотека jsoup (Java HTML Parser). Jsoup является библиотекой Java для работы с HTML. Jsoup обеспечивает удобный API для извлечения и обработки данных, используя лучшие методы DOM, CSS и JQuery.

Jsoup реализует спецификацию WHATWG HTML5 и анализирует HTML с тем же DOM, как это делают современные браузеры. Библиотека позволяет очищать и разбирать HTML из URL, файл или строку, находить и извлекать данные, используя DOM или CSS селекторы, манипулировать с HTML элементами, атрибутами и текстом. Результатом программ является текстовый формат, с которым легко работать в Microsoft Excel для построения гистограмм. Одним из входных параметров программы является число столбцов результирующей гистограммы.

1.3. Пример разработанной компьютерной программы

Большой объем экспериментальных данных требуют значительного времени для анализа расположения участков хромосомных контактов в геноме относительно генов, концов хромосом, аннотированных мРНК, повторов и другой геномной информации.

В ходе работы была разработана компьютерная программа с графическим пользовательским интерфейсом для статистического анализа расположения генов относительно хромосомных петель и распределения сайтов связывания относительно структурной информации о генах.

Пользователь может загрузить свой текстовый файл с генами. При нажатии на кнопку «Выбрать» появляется диалоговое окно выбора файла. Есть возможность автоматической загрузки данных генома из Genome Browser UCSC RefSeq для мыши и человека (последние версии данных на 2015 г.). В случае, если пользователь предпочел

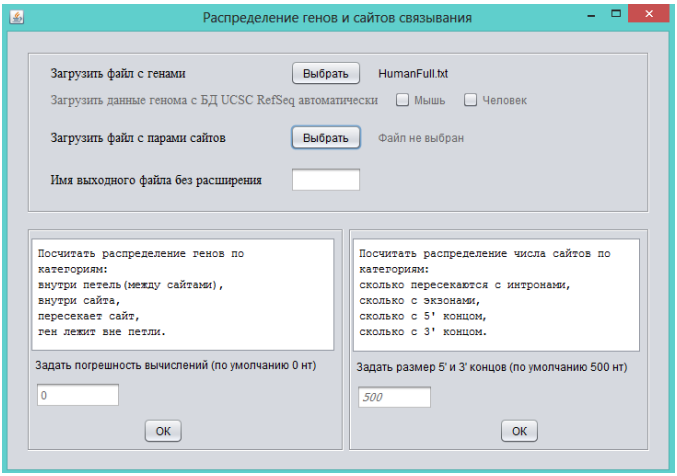

Рис. 2. Пример главного окна разработанной программы

загрузку своего файла, меню автоматической загрузки имеет серый цвет (как неактивное), меняя цветом индикацию активности окна (рис. 2).

В нижней части главного окна (рис. 2) показаны два функционала. Слева — функционал расчета распределения генов относительно петель сайтов связывания по категориям:

- ген лежит внутри петли («не задевая» сайты),
- ген пересекает сайт,
- ген лежит за пределами петли.

Для вычисления результата пользователь может задать погрешность в парах нуклеотидов в соответствующем окне (go умолчанию это значение равно 0). Справа в меню — представлена возможность посчитать распределение сайтов относительно структурной информации о генах, а именно: экзоны, интроны, 5’ и 3’ концы.

В случае, если пользователь не выбрал все необходимые входные данные программы и нажал «OK», появляются стандартные диалоговые окна, сообщающие об ошибке (рис. 3).

Таким образом, программный комплекс отвечает стандартным требованиям контроля входных данных.

После того, как пользователь заполнил все поля верно, перед ним появляется таблица, в которой он должен указать заголовки к

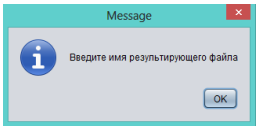

Рис. 3. Проверка наличия входных данных

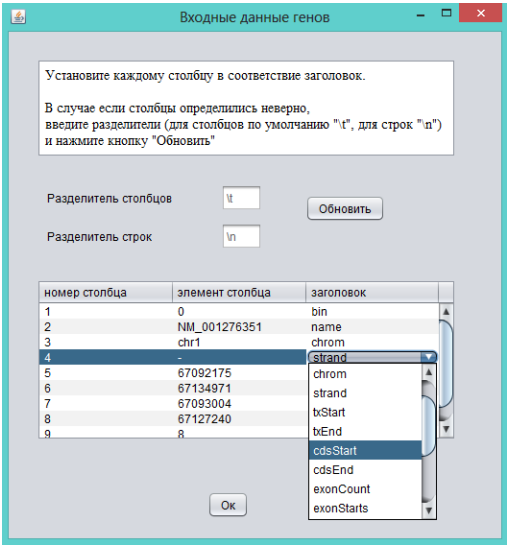

Рис. 4. Таблица для редактирования входных данных

столбцам в его входных данных (в случае если они не совпадают со стандартными заголовками базы данных RefSeq UCSC Genome Browser).

При нажатии на кнопку «OK» программа начинает вычислительные работы. Результат записывается в выходной файл с именем, которое пользователь задавал в первом окне.

На рис. 4 представлен интерфейс для редактирования входных данных в табличном формате на примере стандарта аннотации генов в базе данных RefSeq.

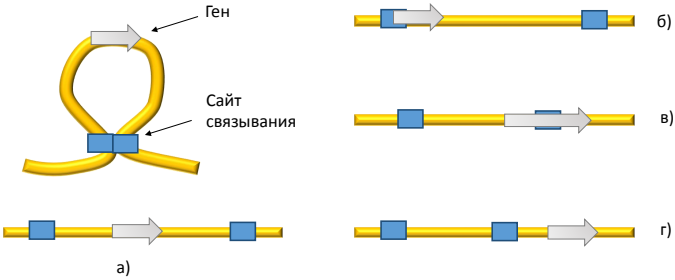

Рис. 5. Модель расположения генов относительно петель сайтов связывания

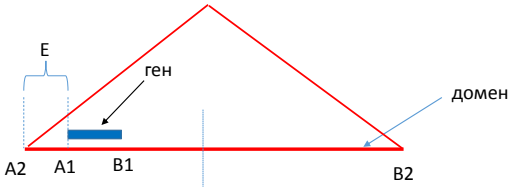

Рис. 6. Определение гена «на границе» пространственного топологического домена

2. Анализ расположения генов относительно участков хромосомных контактов

С помощью компьютерного анализа данных хромосомных контактов можно выделять группы генов, которые совместно экспрессируются, либо имеют схожие функциональные характеристики для последующей реконструкции генных сетей.

За основу подсчета распределения взята следующая модель расположения генов относительно сайтов связывания (рис. 5).

Особое внимание уделялось генам, попавшим на так называемые границы доменов (рис. 6). Погрешность вычислений E (определение «границы») можно изменять.

Полученные списки генов, расположенных в топологических доменах по данным Hi-C для двух типов клеток — сперматозоиды и фибробласты мыши, были в дальнейшем проанализированы при помощи доступных интернет ресурсов STRING (реконструкция генных сетей) и DAVID (категории генных онтологий).

3. Исследование генов в топологических хромосомных доменах

3.1. Реконструкция генных сетей для заданных наборов генов

Рассматривались биологические функции выделенных генов в структуре генных сетей. Был проведен статистический анализ структуры генной сети, образованной генами, выделенными на границах топологических доменов мыши. Предполагается, что хромосомные контакты (физические молекулярные контакты хромосом в ядре клетки) для таких генов приводят к функциональным связям, которые можно найти при анализе баз данных. С помощью разработанной программы проведено исследование аннотированных ранее генных сетей. В частности, были проанализированы генные сети циркадного ритма и регуляции холестерина по расположению на хромосомах мыши.

Гены, контактирующие своими участками в трехмерном расположении хромосом, имеют больше контактов в сети (как белок-белковых контактов, так и регуляторных взаимодействий). Было подсчитано, что общее число генов на границах доменов у фибробластов составляет 698. Из них 88 генов образуют 160 пар связей, что составляет 12% генов от общего числа расположенных на границах пространственных доменов. Общее число таких генов у сперматозоидов составляет 314. Однако только 13 генов участвуют в 10 парах связей, что составляет 4% от общего числа генов, лежащих на границах пространственных доменов.

Показано, что большее число генов лежит в доменах меньшего размера (от 300 тыс. нт до 1600 тыс. нт).

Проанализированы списки генов, находящихся на границах пространственных доменов, на связи коэкспрессии и общие категории генных онтологий (рис. 7). Полученный результат был отсортирован по числу генов, имеющих общие категории генных онтологий. Наибольшее число генов из такого списка отвечают за белки фосфопротеины. Однако наиболее значимые категории онтологий были связаны с функциями мембраны клетки.

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).

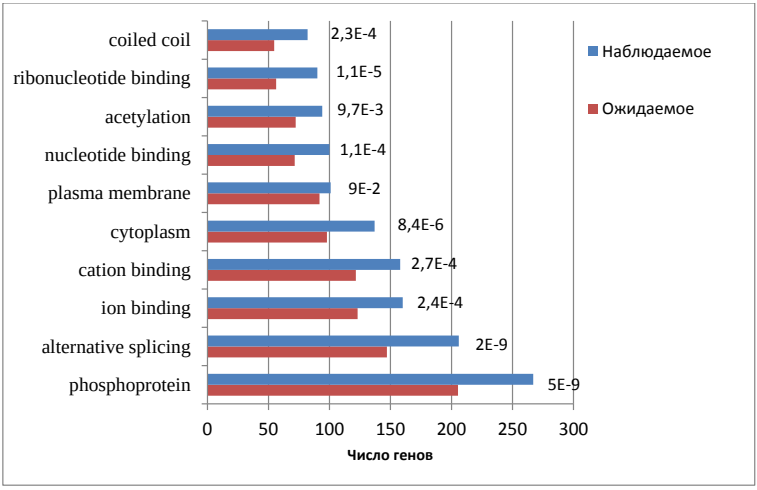

Рис. 7. Категории генных онтологий, расположенных на границах доменов в клетках фибробластов

Генные сети млекопитающих, аннотированные в ИЦиГ СО РАН по клеточному циклу (циркадному ритму) и метаболизму холестерина в клетке, были проанализированы на расположение их генов в пространственных доменах в геноме мыши (для анализа рассматривались гены мыши). Результаты показали, что 8 пар генов располагаются рядом на доменах клеток типа фибробласты, что составляет 14% от общего числа генов, входящих в сеть циркадного ритма мыши. На клетках типа сперматозоид 13 пар генов, что составляет 23%. Это достаточно высокий показатель связности генной сети, свидетельствующий о фундаментальности процессов, контролируемых генами в этой генной сети. Генная сеть метаболизма холестерина на обоих типах клеток дает только одну пару генов, что свидетельствует об отсутствии ее функциональности в таких клетках, относящихся к раннему развитию.

3.2. Развитие программ анализа данных от ChIP-seq к ChIA-PET

Для изучения генома человека с помощью данных экспрессии был создан набор компьютерных программ для работы с существующими базами данных. Кроме работы с доменами, рассмотрено расположения

генов относительно хромосомных петель, образованных парами сайтов CTCF в геноме человека.

Выделены списки генов, пересекающиеся с сайтами связывания транскрипционного фактора CTCF, построены их распределения по хромосомам. По полученным результирующим диаграммам можно судить, что расположение генов относительно хромосомных петель, не случайно. Большая часть генов лежит внутри петель сайтов связывания и не взаимодействует с нуклеотидными координатами сайтов. По объединенному списку, состоящего из генов, которые пересекают сайт CTCF, и генов, один из краев которых попадает в координаты сайта, было построено распределение относительно внутренней структуры гена.

В ходе работы создано и протестировано программное обеспечение для локализации мотивов в нуклеотидной последовательности, заданной в FASTA формате, а также в формате геномных координат (внутри пика ChIP-seq) и обладающее дополнительными функциями для преобразования форматов и визуализации результатов. Код находится под свободной лицензией GNU GPL v2.0.

Используя собственное программное обеспечение, был выполнен поиск точной локализации сайтов 13 транскрипционных факторов (Oct4, Sox2, c-Myc, CTCF, E2f1, Esrrb, Klf4, Nanog, n-Myc, Smad1, STAT3, Tcfcp2i1, Zfx) по позиционным матрицам частот для каждого фактора в участках связывания, ранее с ограниченной точностью определенных с помощью ChIP-seq в геноме мыши. Экспериментально установленное число сайтов варьировано от нескольких тысяч до десятков тысяч. Процент присутствия мотивов в пиках ChIP-seq составил от 25% до 99%. Рассмотрены данные о нуклеотидных мотивах сайтов связывания, для которых доступны данные ChIA-PET, выполнено уточнение мотивов CTCF.

Заключение

Вычислительная сложность анализа данных о регуляторных районах генов увеличивается при рассмотрении взаимодействий между генами в пространстве ядра клетки (трехмерные взаимодействия), что требует разработки новых программных средств и адаптации программных конвейеров. Решение задач обработки объемных данных ChIP-seq и смежных технологий ChIA-PET и Hi-C необходимо для обеспечения технической поддержки геномных исследований [**5**,**12**,**18**].

Ранее подобные технические средства были реализованы в программном комплексе ICGenomics [2], представленном на ССКЦ СО РАН (ЦКП «Биоинформатика» СО РАН).

В результате работы был развит набор программ и утилит для обработки данных ChIP-seq. Программы были написаны на языках Java, C++, Python. Разработан текстовый и графический пользовательский интерфейсы, предусмотрена возможность автоматической загрузки референсного генома с базы данных UCSC Genome Browser и возможность загружать данные любого геномного формата (bed-файлы, координаты, участки последовательностей).

По результатам анализа выделены категории генных онтологий для полученных по данным ChIP-технологий групп генов, контактирующих на хромосомах. Выполнен поиск точной локализации сайтов 13 транскрипционных факторов, уточнены их нуклеотидные мотивы и контекстное окружение (сложность текста). Продолжается интеграция программных инструментов с расчетом экспрессионных данных RNA-seq и применениям для геномов модельных организмов, полученным в ИЦиГ СО РАН [18].

Благодарности. *Авторы благодарны Н. Р. Баттулину, Л. О. Брызгалову, Н. Л. Подколодному, а также ССКЦ СО РАН за предоставление данных и поддержку работы.*

Список литературы

[1] X. Chen, H. Xu, P. Yuan et al.. “Integration of external signaling pathways with the core transcriptional network in embryonic stem cells”, *Cell*, **133**:6 (2008), pp. 1106–1117, URL http://www.ncbi.nlm.nih.gov/pubmed/18555785 ↑ 129, 130, 134.

[2] 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 ↑ 130, 142.

[3] 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, URL http://vavilov.elpub.ru/index.php/jour/article/view/240 ↑ 130, 131, 134.

[4] А. М. Спицина, Ю. Л. Орлов, Н. Н. Подколодная и др.. «Суперкомпьютерный анализ геномных и транскриптомных данных, полученных с помощью технологий высокопроизводительного секвенирования

ДНК», *Программные системы: теория и приложения*, **6**:1(23) (2015), с. 157–174, URL http://psta.psiras.ru/read/psta2015_1_157-174.pdf ↑ 130.

[5] 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), pp. 211, URL http://www.ncbi.nlm.nih.gov/pubmed/22987856 ↑ 130, 141.

[6] 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://www.bionet.nsc.ru/vogis/download/17-4(2)/09_Medvedeva.pdf ↑ 130.

[7] 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 ↑ 130.

[8] F. Ay, T. L. Bailey, W. S. Noble. “Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts”, *Genome Res.*, **24**:6 (2014), pp. 999–1011, URL http://www.ncbi.nlm.nih.gov/pubmed/24501021 ↑ 131.

[9] M. J. Fullwood, M. H. Liu, Y. F. Pan et al.. “An oestrogen-receptor-alpha-bound human chromatin interactome”, *Nature*, **462**:7269 (2009), pp. 58–64, URL http://www.ncbi.nlm.nih.gov/pubmed/19890323 ↑ 131.

[10] 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), pp. 84–98, URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3339270/ ↑ 131.

[11] Н. Р. Баттулин, В. С. Фишман, Ю. Л. Орлов, А. Г. Мензоров, Д. А. Афонников, О. Л. Серова. «3С-методы в исследованиях пространственной организации генома», *Вавиловский журнал генетики и селекции*, **16**:4/2 (2012), с. 872–876, URL http://www.nsu.ru/xmlui/bitstream/handle/nsu/4125/2014_V12_No2_11.pdf ↑ 132.

[12] G. Li, M. J. Fullwood, H. Xu et al.. “ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing”, *Genome Biol.*, **11**:2 (2010), pp. R22, URL http://genomebiology.com/content/11/2/R22 ↑ 133, 141.

[13] C. Li, X. Dong, H. Fan et al.. “The 3DGD: a database of genome 3D structure”, *Bioinformatics*, **30**:11 (2014), pp. 1640–1642, URL http://www.ncbi.nlm.nih.gov/pubmed/24526713 ↑ 133.

[14] 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), R130, URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3091323/ ↑ 133.

[15] Д. А. Полунин, И. А. Штайгер, В. М. Ефимов. «Разработка микрочиповых данных», *Вестник НГУ. Серия: Информационные технологии*, **12**:2 (2014), с. 90–98, URL http://www.nsu.ru/xmlui/handle/nsu/4125 ↑ 133.

[16] В. Н. Бабенко, В. Н. Максимов, Е. В. Кулакова и др.. «Полногеномный анализ пулированных выборок ДНК когорт человека», *Вавиловский журнал генетики и селекции*, **18**:4/2 (2014), с. 847–855, URL http://www.bionet.nsc.ru/vogis/download/18-4-2/003_Babenko.pdf ↑ 133.

[17] Н. Н. Кудрявцева, А. Л. Маркель, Ю. Л. Орлов. «Агрессивное поведение: генетико-физиологические механизмы», *Вавиловский журнал генетики и селекции*, **18**:4/3 (2014), с. 1133–1155, URL http://www.bionet.nsc.ru/vogis/download/18-4-3/11_Kudrjavtseva.pdf ↑ 134.

[18] N. Battulin, V. S. Fishman, A. M. Mazur et al.. “Comparison of the three-dimensional organization of sperm and fibroblast genomes using the Hi-C approach”, *Genome Biol.*, **16**:1 (2015), pp. 77, URL http://www.ncbi.nlm.nih.gov/pubmed/25886366 ↑ 141, 142.

Рекомендовал к публикации *Программный комитет*
Третьего национального суперкомпьютерного форума ***НСКФ-2014***

Об авторах:

Екатерина Викторовна Кулакова
Магистрант НГУ. Область научных интересов: биоинформатика, суперкомпьютерные вычисления.
e-mail: kylakovaekaterina@gmail.com

Анастасия Михайловна Спицина
Магистрант НГУ. Область научных интересов: биоинформатика, суперкомпьютерные вычисления.
e-mail: anastasia.spitsina@gmail.com

Нина Геннадьевна Орлова
Окончила Новосибирский Государственный Университет в 1991 г., кандидат физико-математических наук, старший научный сотрудник НГУ, доцент СибУПК. Область научных интересов: статистика, компьютерная геномика.
e-mail: orlovanina2@mail.ru

Артур Игоревич Дергилев
Студент НГУ. Область научных интересов: биоинформатика, суперкомпьютерные вычисления.
e-mail: arturd1993@yandex.ru

Анатолий Владленович Свичкарев
Студент НГУ. Область научных интересов: биоинформатика, геномика, суперкомпьютерные вычисления.
e-mail: tolik0393@mail.ru

Наталья Сергеевна Сафронова

Магистрант НГУ. Область научных интересов: биоинформатика, дискретная математика.

e-mail: taschasafronova@mail.ru

Игорь Геннадьевич Черных

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

e-mail: chernykh@parbz.sscc.ru

Юрий Львович Орлов

Окончил Новосибирский Государственный Университет в 1991 г., д.б.н., 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

Пример ссылки на эту публикацию:

Е. В. Кулакова, А. М. Спицина, Н. Г. Орлова, А. И. Дергилев и др. «Программы анализа геномных данных секвенирования, полученных на основе технологий ChIP-seq, ChIA-PET и Hi-C», *Программные системы: теория и приложения*, 2015, **6**:2(25), с. 129–148.

URL http://psta.psiras.ru/read/psta2015_2_129-148.pdf

Yekaterina Kulakova, Anastasiya Spitsina, Nina Orlova, Artur Dergilev, Anatoliy Svichkarev, Natal'ya Safronova, Igor' Chernykh, Yuriy Orlov. *Supercomputer analysis of genomics and transcriptomics data revealed by high-throughput DNA sequencing.*

Abstract. Growing volumes of genomic data on transcription factor binding site location, chromosome contacts and the genome features annotations basedon modern sequencing technologies need development of new software and algorithms for processing and analysis of such data. The supercomputing allows to study problems of transcription regulation at qualitatively new level. We consider problems of genome sequencing based on ChIP-seq, ChIA-PET and Hi-C technologies. We present computer programs to solve these tasks and discuss future development. (*In Russian*).

Key Words and Phrases: Bioinformatics, DNA sequencing, Immunoprecipitation, Chromosome contacts, Gene expression regulation, Databases.

References

[1] X. Chen, H. Xu, P. Yuan et al.. "Integration of external signaling pathways with the core transcriptional network in embryonic stem cells", *Cell*, **133**:6 (2008), pp. 1106–1117, URL http://www.ncbi.nlm.nih.gov/pubmed/18555785.

[2] Yu. L. Orlov, A. O. Bragin, I. V. Medvedeva i dr.. "ICGenomics: a program complex for analysis of symbol sequences in genomics", *Vavilov Journal of Genetics and Breeding*, **16**:4/1 (2012), pp. 732–741 (in Russian), URL http://vavilov.elpub.ru/index.php/jour/article/view/70.

[3] Yu. L. Orlov. "Computer-assisted study of the regulation of eukaryotic gene transcription on the base of data on chromatin sequencing and precipitation", *Vavilov Journal of Genetics and Breeding*, **18**:1 (2014), pp. 193–206 (in Russian), URL http://vavilov.elpub.ru/index.php/jour/article/view/240.

[4] A. M. Spitsina, Yu. L. Orlov, N. N. Podkolodnaya i dr.. "Supercomputer analysis of genomics and transcriptomics data revealed by high-throughput DNA sequencing", *Program systems: theory and applications*, **6**:1(23) (2015), pp. 157–174 (in Russian), URL http://psta.psiras.ru/read/psta2015_1_157-174.pdf.

[5] 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), pp. 211, URL http://www.ncbi.nlm.nih.gov/pubmed/22987856.

[6] I. V. Medvedeva, O. V. Vishnevskiy, N. S. Safronova i dr.. "Computer analysis of the data on gene expression in brain cells obtained by microarray tests and high-throughput sequencing", *Russian Journal of Genetics: Applied Research*, **4**:4 (2014), pp. 259–266.

[7] Yu. L. Orlov, V. M. Yefimov, N. G. Orlova. "Statistical estimates of transposable element expression in the human genome based on clinical microarray data on expression", *Vavilov Journal of Genetics and Breeding*, **15**:2 (2011), pp. 327–339 (in Russian), URL http://www.bionet.nsc.ru/vogis/pict_pdf/2011/15_2/12.pdf.

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

[9] M. J. Fullwood, M. H. Liu, Y. F. Pan et al.. “An oestrogen-receptor-alpha-bound human chromatin interactome”, *Nature*, **462**:7269 (2009), pp. 58–64, URL http://www.ncbi.nlm.nih.gov/pubmed/19890323.

[10] 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), pp. 84–98, URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3339270/.

[11] N. R. Battulin, V. S. Fishman, Yu. L. Orlov, A. G. Menzorov, D. A. Afonnikov, O. L. Serova. “3C-based methods for 3D genome organization analysis”, *Vavilov Journal of Genetics and Breeding*, **16**:4/2 (2012), pp. 872–876 (in Russian), URL http://vavilov.elpub.ru/index.php/jour/article/view/85.

[12] G. Li, M. J. Fullwood, H. Xu et al.. “ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing”, *Genome Biol.*, **11**:2 (2010), pp. R22, URL http://genomebiology.com/content/11/2/R22.

[13] C. Li, X. Dong, H. Fan et al.. “The 3DGD: a database of genome 3D structure”, *Bioinformatics*, **30**:11 (2014), pp. 1640–1642, URL http://www.ncbi.nlm.nih.gov/pubmed/24526713.

[14] 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), R130, URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3091323/.

[15] D. A. Polunin, I. A. Shtayger, V. M. Yefimov. “JACOBI 4 software for multivariate analysis of microarray data”, *Vestnik NSU: Information Technologies*, **12**:2 (2014), pp. 90–98 (in Russian), URL http://www.nsu.ru/xmlui/handle/nsu/4125.

[16] V. N. Babenko, V. N. Maksimov, Ye. V. Kulakova i dr.. “Genome-wide snp allelotyping of human cohorts by pooled DNA samples”, *Vavilov Journal of Genetics and Breeding*, **18**:4/2 (2014), pp. 847–855 (in Russian), URL http://www.bionet.nsc.ru/vogis/download/18-4-2/003_Babenko.pdf.

[17] N. N. Kudryavtseva, A. L. Markel’, Yu. L. Orlov. “Aggressive behavior: genetic and physiological mechanisms”, *Vavilov Journal of Genetics and Breeding*, **18**:4/3 (2014), pp. 1133–1155 (in Russian), URL http://www.bionet.nsc.ru/vogis/download/18-4-3/11_Kudrjavtseva.pdf.

[18] N. Battulin, V. S. Fishman, A. M. Mazur et al.. “Comparison of the three-dimensional organization of sperm and fibroblast genomes using the Hi-C approach”, *Genome Biol.*, **16**:1 (2015), pp. 77, URL http://www.ncbi.nlm.nih.gov/pubmed/25886366.

Sample citation of this publication:

Yekaterina Kulakova, Anastasiya Spitsina, Nina Orlova, Artur Dergilev, Anatoliy Svichkarev, Natal’ya Safronova, Igor’ Chernykh, Yuriy Orlov. “Supercomputer analysis of genomics and transcriptomics data revealed by high-throughput DNA sequencing”, *Program systems: theory and applications*, 2015, **6**:2(25), pp. 129–148. (*In Russian.*) URL `http://psta.psiras.ru/read/psta2015_2_129-148.pdf`