ПРОГРАММНЫЙ ПАКЕТ RNAEXPLORER ДЛЯ ПРЕДСКАЗАНИЯ ВАРИАНТОВ АЛЬТЕРНАТИВНОГО СПЛАЙСИНГА В ПЕРВИЧНЫХ МРНК ХИМЕРНОГО ОНКОГЕНА RUNX1/RUNX1T1 ЧЕЛОВЕКА

Яцков Н. Н., Скакун В. В., Гринев В. В.
Кафедра системного анализа и компьютерного моделирования, кафедра генетики, Белорусский государственный университет
Минск, Республика Беларусь
E-mail: {yatskou, skakun}@bsu.by

В работе представлен программный пакет RNAexploreR для предсказания событий альтернативного сплайсинга в первичных мРНК онкогена RUNX1/RUNX1T1 человека. Выполнен анализ набора экспериментальных данных для гена RUNX1/RUNX1T1 с использованием базового и распараллеленного вариантов программного пакета.

Введение

Гибридный онкоген RUNX1-RUNX1T1 является одним из факторов инициации и формирования острого миелоидного лейкоза у детей и взрослых. Данный онкоген кодирует более 130 изоформ РНК, закономерности образования которых до конца не установлены [1]. Современные работы нацелены на выяснение принципов, по которым идет комбинаторика экзонов во время сплайсинга и на разработку алгоритмических и программных средств для анализа и предсказания разнообразных вариантов РНК [2]. Однако интерактивные, доступные для широкого круга пользователей программные приложения, реализующие разработанные алгоритмы, практически отсутствуют или представлены локально в ограниченном виде.

Целью работы является разработка интерактивного высокопроизводительного программного пакета для предсказания вариантов альтернативного сплайсинга в первичных мРНК онкогена RUNX1/RUNX1T1 человека.

I. Программные средства для интеллектуального анализа данных

В настоящее время в открытом доступе предоставлено большое количество программных средств интеллектуального анализа данных, среди которых можно выделить: WEKA, Tanagra, Rapid Miner, KNIME, Python- и R-платформы [3]. Достоинствами того или иного программного ресурса являются: вычислительная производительность, широкий набор подключаемых библиотек, кроссплатформенность, возможность выполнения распараллеленных вычислений и работы напрямую с базами и хранилищами данных.

Основным преимуществом среды статистического программирования R является возможность использования огромного набора биоинформационных алгоритмов, алгоритмов интеллектуального анализа данных, разнообразных статистических вычислительных ресурсов научного сообщества [4, 5]. Главным недостатком является невысокая вычислительная производительность, однако данное ограничение можно частично или полностью устранить с помощью использования процедур распараллеливания вычислений или подключения библиотек высокопроизводительных математических вычислений (например, библиотек Microsoft R Open (MRO) и Intel Math Kernel Library (MKL)). Наиболее популярные пакеты для разработки пользовательских интерфейсов программных приложений, интегрирующие R-коды, это: gWidgets, rpanel, svDialogs, RGtk2, qtbase, tcltk [5]. Новое направление в разработке R-приложений связано с созданием «реактивных» веб-интерфейсов с использованием пакета Shiny и размещением программной реализации на ресурсе shinyapps.io, предоставляемом разработчиками открытого программного обеспечения RStudio. Достоинством данного подхода является возможность удаленной работы с приложением широкой научной аудитории пользователей в режиме on-line через глобальную сеть Internet.

Для реализации программного приложения в работе выбраны вычислительная среда R и пакет Shiny для создания веб-интерфейса разработанного приложения.

II. Распараллеливание вычислений

Научным сообществом разработано большое количество R-пакетов для реализации распараллеленных вычислений, среди которых можно выделить пакеты Rmpi, snow, snowfall, foreach и parallel [5]. Для автоматизации и упрощения работы пользователя на основе пакета snow и модели распараллеливания SOCKETS разработан пакет foreach, позволяющий относительно просто реализовать решаемую задачу в виде независимых циклов вычислений на многоядерном процессоре.

Для реализации в работе выбран пакет foreach, т.к. предоставляет: 1) логическую чет-

ко выстроенную структуру для распараллеливания кода; 2) эффективные автоматизированные средства распараллеливания; 3) возможность распараллеливания на многоядерном процессоре; 4) возможность анализа больших массивов данных, реализуемую с помощью пакета iterators; 5) набор примеров по распараллеливанию задач интеллектуального анализа данных.

III. Описание вычислительного эксперимента

Тестирование пакета произведено на двух вычислительных системах (персональных компьютерах), далее представленных как ПК1 и ПК2, характеризующихся различными вычислительными возможностями и производительностью. ПК1 содержит двухядерный процессор DualCore Intel Pentium E5700, работает под 32 битной OS MS Windows XP. Данный ПК используется как эталонный вариант компьютера. ПК2 является более продвинутой вычислительной системой в смысле распараллеливания решаемых задач, имеет четырехядерный процессор QuadCore Intel Core 2 Quad Q9400, работает под 64 битной OS MS Windows 10. ПК2 является хорошим примером компьютерного (4-х ядерного) кластера для решения задач распараллеливания.

В работе исследуются способы повышения эффективности пакета с помощью подключения высокопроизводительной библиотеки Intel MKL и распараллеливания программного кода.

IV. Результаты

Разработан программный пакет RNAexploreR. On-line версия пакета доступна по ссылке `https://dsa-cm.shinyapps.io/NIR_bio_code_Sh/`. Главное окно интерфейса пакета состоит из набора панелей, соответствующих основным этапам анализа данных. На каждом этапе анализа пользователь должен загрузить требуемый файл данных (файлы экзонов, экспериментально подтвержденных транскриптов, теоретически предсказанных транскриптов) и установить системные параметры алгоритмов интеллектуального анализа данных. Пример результирующего этапа анализа данных представлен на рис. 1.

Выполнено исследование эффективности работы базового и распараллеленного вариантов программного пакета на примере анализа полного набора экспериментальных данных для гена RUNX1/RUNX1T1 [1]. Проводилось от 3 до 6 запусков пакета на каждом из ПК.

В результате распараллеливания алгоритмов удалось существенно сократить время работы пакета на 41% на ПК1 (с 8.66 до 5.13 мин) и на 316% ПК2 (с 6.84 до 1.80 мин). Распараллеленный вариант пакета работает в 5 раз быстрее на ПК2, чем его нерспараллеленная Shiny-версия на ПК1.

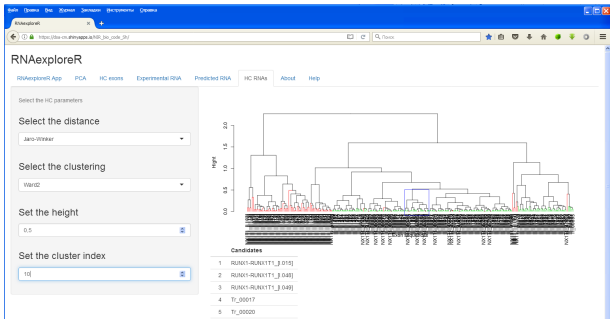

Рис. 1 – Результаты работы алгоритма иерархической кластеризации пула уникальных экспериментальных (красный цвет) и теоретических (зеленый цвет) транскриптов онкогена RUNX1/RUNX1T1. Для демонстрации результаов список теоретически предсказанных транскриптов сокращен до 3000

V. Заключение

Проведено исследование существующих свободных программных средств интеллектуального анализа данных для реализации программного приложения для предсказания событий альтернативного сплайсинга в первичных мРНК химерного онкогена RUNX1/RUNX1T1 человека. Выбраны наиболее оптимальные программные средства для реализации исследуемой задачи – вычислительная среда R и пакет Shiny для создания веб-интерфейса приложения. Разработано веб-приложение и распараллеленный вариант пакета RNAexploreR. Выполнен анализ набора данных для гена RUNX1/RUNX1T1 с использованием базового и распараллеленного вариантов программного пакета на различных компьютерах.

VI. Список литературы

1. Ильюшёнок, И. Н. Изучение закономерностей сплайсинга РНК гибридного онкогена RUNX1-RUNX1T1 человека с помощью методов интеллектуального анализа данных и высокопроизводительного секвенирования / И. Н. Ильюшёнок, Е. П. Гунько, М. Л. Антонович, Н. Н. Яцков, А. М. Кустанович, А. Ю. Сухаревский, В. В. Скакун, В. В. Гринев // Молекулярная и прикладная генетика. – 2017. – Том 23. – С. 92–101.
2. Grinev, V. V. Decoding of exon splicing patterns in the human RUNX1-RUNX1T1 fusion gene / V. V. Grinev, A. A. Migas, A. D. Kirsanava, O. A. Mishkova, N. Siomava, T. V. Ramanouskaya, A. V. Vaitsiankova, I. M. Ilyushonak, P. V. Nazarov, L. Vallar, O. V. Aleinikova // Int. J. Biochem. Cell Biol. – 2015. – Vol. 68. – P. 48–58.
3. Kdnuggets [Electronic recourse] / – Mode of access: https://www.kdnuggets.com. – Date of access: 25.08.2018.
4. Bioconductor [Electronic recourse] / – Mode of access: http://www.bioconductor.org. – Date of access: 25.08.2018.
5. The R Project for Statistical Computing [Electronic recourse] / – Mode of access: http://www.r-project.org. – Date of access: 25.08.2018.