Министерство образования Республики Беларусь
Учреждение образования
«Полоцкий государственный университет»

ИНФОРМАЦИОННО-КОММУНИКАЦИОННЫЕ ТЕХНОЛОГИИ: ДОСТИЖЕНИЯ, ПРОБЛЕМЫ, ИННОВАЦИИ (ИКТ-2018)

Электронный сборник статей
I Международной научно-практической конференции,
посвященной 50-летию Полоцкого государственного университета

(Новополоцк, 14–15 июня 2018 г.)

Новополоцк
Полоцкий государственный университет
2018

Информационно-коммуникационные технологии: достижения, проблемы, инновации (ИКТ-2018) [Электронный ресурс] : электронный сборник статей I международной научно-практической конференции, посвященной 50-летию Полоцкого государственного университета, Новополоцк, 14–15 июня 2018 г. / Полоцкий государственный университет. – Новополоцк, 2018. – 1 электрон. опт. диск (CD-ROM).

Представлены результаты новейших научных исследований, в области информационно-коммуникационных и интернет-технологий, а именно: методы и технологии математического и имитационного моделирования систем; автоматизация и управление производственными процессами; программная инженерия; тестирование и верификация программ; обработка сигналов, изображений и видео; защита информации и технологии информационной безопасности; электронный маркетинг; проблемы и инновационные технологии подготовки специалистов в данной области.

Сборник включен в Государственный регистр информационного ресурса. Регистрационное свидетельство № 3201815009 от 28.03.2018.

Компьютерный дизайн М. Э. Дистанова.

Технические редакторы: Т. А. Дарьянова, О. П. Михайлова.

Компьютерная верстка Д. М. Севастьяновой.

211440, ул. Блохина, 29, г. Новополоцк, Беларусь
тел. 8 (0214) 53-21-23, e-mail: irina.psu@gmail.com

ISBN 978-985-531-602-3

УДК 51-76[577.21+004.4]; 577.21:519.1

ВЕБ-ПРИЛОЖЕНИЕ ДЛЯ ПРЕДСКАЗАНИЯ СОБЫТИЙ АЛЬТЕРНАТИВНОГО СПЛАЙСИНГА В ПЕРВИЧНЫХ мРНК ОНКОГЕНОВ ЧЕЛОВЕКА

канд. физ.-мат. наук Н.Н. ЯЦКОВ, канд. биол. наук В.В. ГРИНЕВ, канд. физ.-мат. наук В.В. СКАКУН
(Белорусский государственный университет, Минск)

Введение. Конститутивный и альтернативный сплайсинг является фундаментальным процессом, протекающим во всех без исключения клетках эукариот и приводящим к образованию зрелых функциональных РНК-продуктов [1]. Однако, несмотря на столь высокую значимость и почти 40-летнюю историю изучения данного процесса, принципы (правила) комбинаторики экзонов во время сплайсинга до сих пор не установлены [2]. Следует отметить об ограниченном применении или даже полном отсутствии стандартов или единых систематизированных статистических подходов к анализу и интерпретации возможных экзонных последовательностей генов человека. 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 [3,4]. 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 веб-приложения для предсказания событий альтернативного сплайсинга в первичных мРНК онкогенов человека. В работе представлен программный пакет в виде веб-приложения Shiny для предсказания событий альтернативного сплайсинга в первичных мРНК онкогенов человека. Для всестороннего исследования разработанного программного пакета используется набор экспериментальных данных, полученных в ходе анализа онкогена RUNX1/RUNX1T1 [3].

Методология. *Объект и предмет исследования.* Объект исследования – альтернативный сплайсинг мРНК онкогена человека. В качестве примера выбран гибридный онкоген RUNX1/RUNX1T1 [3]. Предметом исследования являются закономерности комбинаторики экзонов во время сплайсинга первичных РНК-продуктов онкогена RUNX1/RUNX1T1, экспрессирующихся в лейкозных клетках острого миелоидного лейкоза с транслокацией t(8;21)(q22;q22).

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 [5]. 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 [6-10]. Главным недостатком является невысокая вычислительная производительность, однако данное ограничение можно частично или полностью устранить с помощью использования процедур распараллеливания вычислений и подключения библиотек высокопроизводительных математических вычислений (например, библиотек Microsoft R Open и Intel Math Kernel) [11, 12]. Наиболее популярные пакеты для разработки пользовательских интерфейсов программных приложений, интегрирующие R-коды, это: gWidgets, rpanel, svDialogs, RGtk2, qtbase, tcltk [13]. Новое направление в разработке R-приложений связано с созданием «реактивных» веб-интерфейсов с использованием пакета Shiny [14] 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.

Методика предсказания событий альтернативного сплайсинга. Разработанные алгоритмы запрограммированы на языке R и собраны в единый подход. Основные этапы подхода:

Этап 1. Анализ полного набора признаков экзонов с использованием метода главных компонент [15]. Шкалирование и центрирование данных. Отбор главных компонент, объясняющих 95% вариации в данных.

Этап 2. Иерархическая кластеризация [15] экзонов гена на основе отобранного 95% набора новых признаков (главных компонент). Разбиение экзонов на кластеры и сопоставление каждому кластеру уникального индекса в символах латинского алфавита (от a до z).

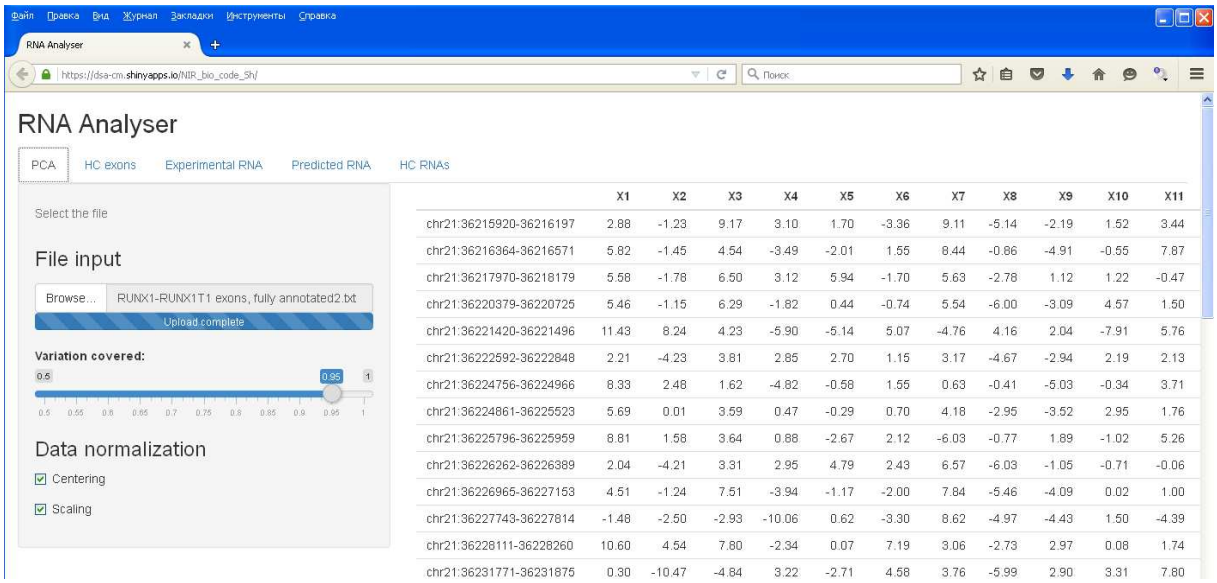
Этап 3. Преобразование символов последовательностей экспериментальных и теоретических транскриптов РНК (от имен экзонов) к меткам кластеров, в которых расположены соответственные экзоны.

Этап 4. Удаление транскриптов дубликатов.

Этап 5. Иерархическая кластеризация пула уникальных (в смысле не дубликатов) экспериментальных и теоретических транскриптов. Представление результатов анализа в виде дендрограммы и списка наиболее вероятных транскриптов.

Результаты. Разработано веб-приложение, интегрирующее реализованные алгоритмы. On-line версия пакета доступна по ссылке https://dsa-cm.shinyapps.io/NIR_bio_code_Sh/. Главное окно интерфейса пакета состоит из пяти панелей, соответствующих пяти этапам анализа. 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 первого и пятого этапов анализа данных для онкогена RUNX1/RUNX1T1 представлен на рисунках 1 и 2. Для демонстрации результатов список теоретически предсказанных транскриптов сокращен до 3000.



	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11
chr21:36215920-36216197	2.88	-1.23	9.17	3.10	1.70	-3.36	9.11	-5.14	-2.19	1.52	3.44
chr21:36216364-36216571	5.82	-1.45	4.54	-3.49	-2.01	1.55	8.44	-0.86	-4.91	-0.55	7.87
chr21:36217970-36218179	5.58	-1.78	6.50	3.12	5.94	-1.70	5.63	-2.78	1.12	1.22	-0.47
chr21:36220379-36220725	5.46	-1.15	6.29	-1.82	0.44	-0.74	5.54	-6.00	-3.09	4.57	1.50
chr21:36221420-36221496	11.43	8.24	4.23	-5.90	-5.14	5.07	-4.76	4.16	2.04	-7.91	5.76
chr21:36222592-36222848	2.21	-4.23	3.81	2.85	2.70	1.15	3.17	-4.67	-2.94	2.19	2.13
chr21:36224756-36224966	8.33	2.48	1.62	-4.82	-0.58	1.55	0.63	-0.41	-5.03	-0.34	3.71
chr21:36224861-36225523	5.69	0.01	3.59	0.47	-0.29	0.70	4.18	-2.95	-3.52	2.95	1.76
chr21:36225796-36225959	8.81	1.58	3.64	0.88	-2.67	2.12	-6.03	-0.77	1.89	-1.02	5.26
chr21:36226262-36226389	2.04	-4.21	3.31	2.95	4.79	2.43	6.57	-6.03	-1.05	-0.71	-0.06
chr21:36226965-36227153	4.51	-1.24	7.51	-3.94	-1.17	-2.00	7.84	-5.46	-4.09	0.02	1.00
chr21:36227743-36227814	-1.48	-2.50	-2.93	-10.06	0.62	-3.30	8.62	-4.97	-4.43	1.50	-4.39
chr21:36228111-36228260	10.60	4.54	7.80	-2.34	0.07	7.19	3.06	-2.73	2.97	0.08	1.74
chr21:36231771-36231875	0.30	-10.47	-4.84	3.22	-2.71	4.58	3.76	-5.99	2.90	3.31	7.80

Рисунок 1 – Таблица экзонов изучаемого онкогена в терминах первых главных компонент (относительная доля разброса, приходящаяся на первые компоненты – 0.95), полученных в ходе анализа 99 экзонов и набора 1438 нормированных признаков

	Candidates
1	RUNX1-RUNX1T1_[I.015]
2	RUNX1-RUNX1T1_[I.048]
3	RUNX1-RUNX1T1_[I.049]
4	Tr_00017
5	Tr_00020
6	Tr_00047

Рисунок 2 – Результаты работы алгоритма иерархической кластеризации пула уникальных экспериментальных (красный цвет) и теоретических (зеленый цвет) транскриптов онкогена RUNX1/RUNX1T1

Заключение. Проведено исследование существующих свободных программных средств интеллектуального анализа данных для реализации программного приложения предсказания событий альтернативного сплайсинга в первичных мРНК онкогенов человека. Выбраны наиболее оптимальные программные средства для реализации исследуемой задачи – вычислительная среда R и пакет Shiny для создания веб-

интерфейса приложения. Разработано веб-приложение, интегрирующее реализованные алгоритмы. Выполнен анализ набора данных для гена RUNX1/RUNX1T1 с использованием разработанного пакета. В дальнейшем планируется улучшение пакета, включающее распараллеливание алгоритмов разработанного подхода, модернизацию и ускорение производительности вычислений за счет оптимизации алгоритмов и подключения высокопроизводительных математических библиотек.

Литература

1. Hang J. [et al.] Shi Y. Structural basis of pre-mRNA splicing. // Science. – 2015. – Vol. 349, № 6253. – P. 1191-1198.
2. Deciphering the splicing code / Y. Barash [et al.] // Nature. – 2010. – Vol. 465, № 729. – P. 53–59.
3. Decoding of exon splicing patterns in the human RUNX1-RUNX1T1 fusion gene / V.V. Grinev [et al.] // Int. J. Biochem. Cell Biol. – 2015. – Vol. 68 – P. 48-58.
4. Изучение закономерностей сплайсинга РНК гибридного онкогена RUNX1-RUNX1T1 человека с помощью методов интеллектуального анализа данных и высокопроизводительного секвенирования / И.Н. Ильюшёнок [и др.] // Молекулярная и прикладная генетика. – 2017. – Т. 23. – С. 92–101.
5. Электронный ресурс. – Режим доступа: https://www.kdnuggets.com.
6. Электронный ресурс. – Режим доступа: http://www.bioconductor.org.
7. Электронный ресурс. – Режим доступа: http://www.r-project.org.
8. Электронный ресурс. – Режим доступа: http://cran.r-project.org.
9. Электронный ресурс. – Режим доступа: http://r-forge.r-project.org.
10. Электронный ресурс. – Режим доступа: https://github.com.
11. Электронный ресурс. – Режим доступа: https://CRAN.R-project.org/view=HighPerformanceComputing.
12. Электронный ресурс. – Режим доступа: https://mran.microsoft.com.
13. Lawrence, M.F. Programming Graphical User Interfaces in R. / M.F. Lawrence, J. Verzani // Chapman & Hall/CRC – The R Series // CRC Press, Taylor & Francis Group, LLC. – 2012. – 463 p.
14. EFS: an ensemble feature selection tool implemented as R-package and web-application / U. Neumann [et al.] // BioData Mining. – 2017.
15. Яцков, Н.Н. Интеллектуальный анализ данных : пособие / Н.Н. Яцков. – Минск : БГУ, 2014. – 151 с.