СЕКЦИЯ «БИОИНФОРМАТИКА»

ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ БОЛЬШИХ ТРАНСКРИПТОМНЫХ ДАННЫХ

В. В. Гринев, Н. Н. Яцков, В. В. Скакун

Белорусский государственный университет, Минск, Беларусь
E-mail: grinev_vv@bsu.by

В статье представлены результаты использования методов интеллектуального анализа данных для решения задач транскриптомики. Обсуждается потенциал таких методов в идентификации сайтов генетического полиморфизма, влияющих на транскрипцию генов, процессинг РНК и структуру кодируемых ими белков. Кроме того, приводятся примеры успешного использования методов интеллектуального анализа данных при изучении экспресии генов, эпигенетического контроля этого процесса, дифференциального сплайсинга и функциональной аннотации транскриптомов.

Ключевые слова: *большие транскриптомные данные, интеллектуальный анализ данных, классификация, предиктивные модели.*

Высокопроизводительные методы прочтения нуклеотидных последовательностей все шире используются в диагностике и прогнозировании течения различных заболеваний человека (включая персонализированную медицину), в спортивной медицине, криминалистике, а также сельском, лесном и охотничьем хозяйствах. При этом все более отчетливо проявляются не проблемы получения первичных данных секвенирования, а проблемы хранения, передачи, анализа и интерпретации таких данных, которые эффективно могут решаться только через тесную кооперацию специалистов из молекулярной биологии, физики, математики и информатики. Междисциплинарная кооперация привела к появлению нового направления научных исследований – анализа больших данных, - включающего разработку нетривиальных методологических и технологических решений, в том числе алгоритмов искусственного интеллекта, распределенных вычислений и т. д.

В настоящей статье авторы делятся опытом и результативностью использования методов интеллектуального анализа данных в транскриптомных исследованиях. Проведенные исследования охватывают несколько уровней реализации наследственной информации человека – от контроля транскрипции до формирования зрелых молекул РНК. Описанные подходы могут быть использованы как в фундаментальных, так и в прикладных разработках, в частности, при постановке дифференциального диагноза, прогнозе ответа на терапию и решения ряда других задач персонализированной медицины с использованием больших мульти-OMICS данных.

НЕСТАБИЛЬНОСТЬ ГЕНОМА

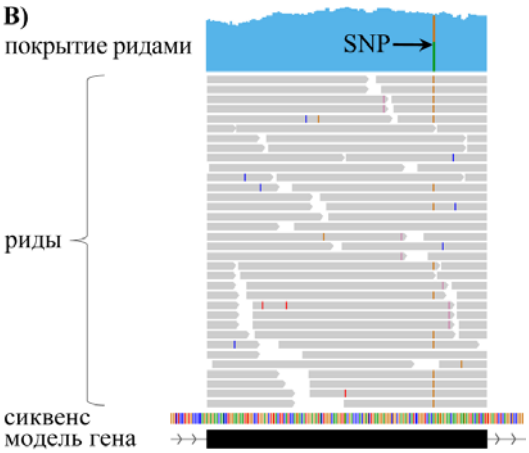
Нуклеотидную последовательность нашего генома можно представить как траекторию в вероятностном пространстве: каждая позиция (сайт) такой последовательности может быть занята одним из четырех нуклеотидов с вероятностью, отличной от 1. При этом есть сайты (рис. 1А), которые очень изменчивы, варианты которых встречаются в популяциях людей с частотой более 1% и именуются сайтами простого нуклеотидного полиморфизма, или SNPs (от англ. simple/small/short nucleotide polymorphisms). Сюда также следует добавить разнообразные мутации, встречающиеся с частотой менее 1%, и мы получим очень высокую изменчивость нашего генома (рис. 1Б), влияющую на транскрипцию генов, процессинг первичных РНК и структуру кодируемых белков.

А)

геном 1	AGCCAGTTTGACAGCTTG
геном 2	AGCCAGTTTGACTGCTTG
геном 3	AGCCAGTTTGACTGCTTG
геном 4	AGCCAGTTTGACAGCTTG
эталон	AGCCAGTTTGACTGCTTG

Б) Разнообразие сайтов полиморфизма и мутаций в геноме человека

База данных	Последний релиз	Уникальные сайты
ClinVar	19.03.2022 г.	1 201 483
dbSNP	07.03.2022 г.	660 773 127
dbVar	07.03.2022 г.	7 446 975



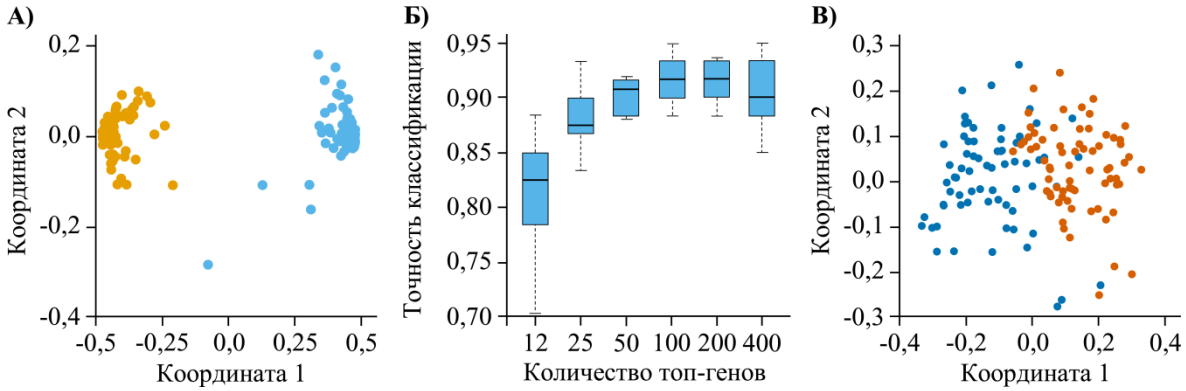
А) Пример полиморфного сайта в геноме человека (сопоставлены индивидуальные геномы и эталонная последовательность).
Б) Разнообразие сайтов полиморфизма и мутаций в геноме человека.
В) Идентификация сайтов полиморфизма с помощью геномного секвенирования (фрагмент гена *WWP2* человека, полиморфный сайт помечен как SNP).

Рис. 1. Генетический полиморфизм человека

Поскольку для многих полиморфных сайтов нашего генома обнаружены ассоциации с фенотипом (физические и умственные способности человека, предрасположенность к заболеваниям, характер ответа на лечение, продолжительность жизни и т. д.), то в настоящее время особое внимание уделяется идентификации таких сайтов в индивидуальных геномах по данным полногеномного или экзомного секвенирования (рис. 1В). Такого рода задача имеет как классическое решение, основанное, например, на точном тесте Фишера (реализовано в бета-версии нашего пакета GSVCalleR [1]), так и нетривиальное, в частности, с использованием нейронных сетей глубокого обучения, что разрабатывается нами в настоящее время.

ДИФФЕРЕНЦИАЛЬНАЯ ЭКСПРЕССИЯ ГЕНОВ И ЕЕ КОНТРОЛЬ

Идентификация генов, экспрессия которых различается в разных типах клеток, на разных этапах развития или при разных условиях, является типовой задачей, решаемой как в фундаментальных, так и в прикладных исследованиях. Дополнение стандартных пайплайнов оценки дифференциальной экспресии генов методами интеллектуального анализа расширяет возможности по работе с транскриптомными данными. Так, алгоритм случайного леса, используя различия в экспресии генов, позволяет надежно диагностировать тип лейкоза у человека (рис. 2А).



А) Многомерное шкалирование матрицы близости между образцами острого лимфобластного (●) и острого миелоидного (●) лейкозов, рассчитанной с помощью алгоритма случайного леса. Точность классификации составила 99,8 % ± 0,5 %.

Б) Влияние количества топовых генов на точность классификации клеток по чувствительности к нок-ауту гена *MSL1*. Значимость генов для процедуры классификации в алгоритме случайного леса рассчитывалась с помощью индекса Джини.

В) Многомерное шкалирование матрицы близости между образцами чувствительных (●) и устойчивых (●) к нок-ауту гена *MSL1* клеток, рассчитанной с помощью алгоритма случайного леса. Точность классификации составила 92,0 % ± 3,1 %.

Рис. 2. Использование алгоритма случайного леса для классификации разных типов клеток человека

Кроме того, гены могут быть ранжированы по их вкладу в точность классификации, что позволяет отобрать только наиболее значимые из них. Так, с помощью индекса Джини мы идентифицировали гены, наиболее значимые в определении чувствительности клеток человека к нок-ауту эпигенетического регулятора транскрипции *MSL1* (рис. 2Б) [2]. По экспресии таких генов возможна дальнейшая классификация разных типов клеток по чувствительности к нок-ауту гена *MSL1* (рис. 2В). Эти же гены могут выступать в качестве приоритетных мишеней при разработке новых молекулярных терапевтиков.

ДИФФЕРЕНЦИАЛЬНЫЙ СПЛАЙСИНГ МОЛЕКУЛ ПЕРВИЧНЫХ РНК И ЕГО КОНТРОЛЬ

Не менее значимым является изучение с помощью методов мультивариантного анализа многомерных данных, касающихся процессинга (в частности, сплайсинга) первичных РНК человека. Так, с помощью метода независимых компонент могут быть идентифицированы такие паттерны сплайсинга, которые позволяют надежно разделить лейкозные клетки и нормальные клетки крови человека, а также разные типы лейкозных клеток, что можно использовать в дифференциальной диагностике заболеваний человека, особенно в персонализированной медицине [3]. Кроме того, с помощью мета-классификаторов удается идентифицировать ключевые регуляторы альтернативного и дифференциального сплайсинга, что имеет как фундаментальную, так и прикладную значимость [4].

ФУНКЦИОНАЛЬНАЯ АННОТАЦИЯ ТРАНСКРИПТОМА

Очень важной компонентой в работе с транскриптомными данными является функциональная аннотация транскриптомов. Тут возможно использование разных подходов, что определяется, главным образом, конечной целью исследования. Так, нами разработан эффективный пайплайн по идентификации открытых рамок считывания в полноразмерных молекулах РНК человека [5], в котором используется векторизация признаков РНК и алгоритм случайного леса. Результаты такой работы позволяют классифицировать РНК на кодирующие и некодирующие, что, опять же, имеет как фундаментальную, так и прикладную значимость.

БИБЛИОГРАФИЧЕСКИЕ ССЫЛКИ

1. Tirtakusuma R., Szoltysek K., Milne P., Grinev V. V. et al. Epigenetic regulator genes direct the fate of multipotent progenitor cell of origin in lineage switched MLL/AF4 leukaemia // Blood (preprint in bioRxiv). 2022. DOI: 10.1101/2021.07.16.452676
2. Radzisheuskaya A., Shliaha P. V., Grinev V. V. et al. Complex-dependent histone acetyltransferase activity of KAT8 determines its role in transcriptional regulation and cellular homeostasis // Molecular Cell. 2021. Vol. 81. P. 1749-1765. DOI: 10.1016/j.molcel.2021.02.012
3. Grinev V. V., Barneh F., Ilyushonak I. M. et al. RUNX1/RUNX1T1 mediates alternative splicing and reorganizes the transcriptional landscape in leukemia // Nature Communications. 2021. Vol. 12. DOI: 10.1038/s41467-020-20848-z
4. Grinev V. V., Migas A. A., Kirsanava A. D. et al. Decoding of exon splicing patterns in the human RUNX1-RUNX1T1 fusion gene // The International Journal of Biochemistry and Cell Biology. 2015. Vol. 68. P. 48-58. DOI: 10.1016/j.biocel.2015.08.017
5. Grinev V. V., Yatskou M. M., Skakun V. V. et al. ORFhunteR: an accurate approach for the automatic identification and annotation of open reading frames in human mRNA molecules // Software Impacts. 2022. In press. DOI: 10.1016/j.simpa.2022.100268