Разработка вычислительного подхода для автоматического определения открытых рамок считывания в молекулах РНК человека

Н.Н. Яцков, В.В. Скакун, В.В. Гринев

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

В настоящее время разработан ряд подходов, позволяющих по данным полнотранскриптомного секвенирования [1, 2] не только восстановить (собрать) структуру всех молекул РНК, присутствующих в клетке, но и дать им количественную, а также качественную характеристику. Так, после сборки кодирующий потенциал молекул РНК может быть оценен с помощью алгоритмов NCBI ORFfinder [3] или CPC2 [4]. К сожалению, ни один из этих инструментов не позволяет сделать обоснованный выбор одной из открытых рамок считывания (ОРС) в случае множественности таковых в изучаемой молекуле РНК. Кроме того, такие алгоритмы работают, как правило, не с целыми транскриптомами, а с индивидуальными молекулами РНК, причем без интеграции с другими приложениями для анализа множества структурно-функциональных характеристик РНК.

В работе предложен вычислительный подход для автоматического определения ОРС в молекулах РНК на основе процедур векторизации нуклеотидных последовательностей и использования классификатора случайного леса.

Методология. Вычислительный подход включает алгоритмы векторизации [5] и случайного леса [6]. Векторизация последовательностей произведена в 104 признака (частоты моно-, ди- и тринуклеотидов [5], параметры модели Bao [7], корреляционные факторы нуклеотидов [8], длинны последовательностей). Этапы анализа.

1. Формирование наборов данных для обучения, представляющих классы истинных (кодирующих) и псевдо (некодирующих) ОРС-кандидатов.
2. Векторизация фрагментов нуклеотидных последовательностей молекул в 104 признака.
3. Обучение метода случайного леса на эталонном наборе данных. Оценка точности (ошибки) классификации на тестируемом наборе данных. Экспорт классификационной модели для определения ОРС молекул РНК.

4. Анализ исследуемых молекул РНК с целью точного определения ОРС: i) определение всевозможных ОРС-кандидатов в молекуле; ii) точное нахождение ОРС с использованием классификационной модели п.3.

Данные. В ходе анализа рассмотрены 4235 некодирующих молекул РНК, не имеющих ОРС, и 113063 кодирующих молекул РНК из базы данных NCBI RefSeq. Класс псевдо ОРС-последовательностей содержит 109230 нуклеотидных фрагмента, полученных из 4235 некодирующих молекул РНК. Класс истинных ОРС-последовательностей включает 108654 реальных ОРС из молекул РНК. Оценка точности определения ОРС произведения на полном наборе кодирующих молекул РНК. Для оценки точности определения ОРС используются координаты ОРС молекул, представленные в базе данных NCBI RefSeq.

Результаты. Вычислительные алгоритмы реализованы на языках программирования R и C++ с использованием открытых библиотек R-функций проектов Bioconductor и CRAN. Анализ данных выполнен на вычислительном сервере, основные характеристики которого – 12 ядерный процессор Intel i9 (3.9 GHz), 64 Gb RAM, 8 Tb HDD. Время вычислений – 10 часов.

Визуализация результатов векторизации ОРС-последовательностей с использованием метода главных компонент [9] представлена на рисунке. Два класса ложных и истинных ОРС последовательностей разделяются.

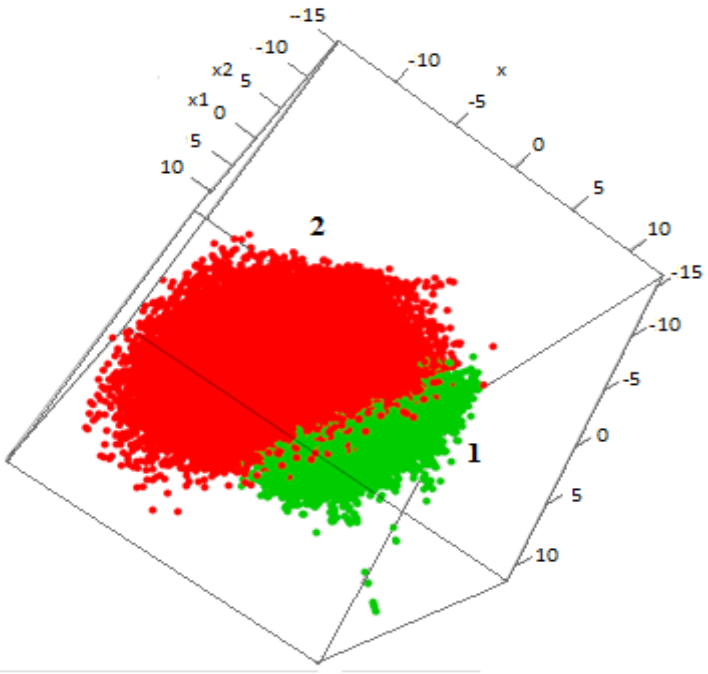

Рис. Результаты применения метода главных компонент к векторизованному набору данных: ОРС-кандидаты кодирующих (1) и некодирующих (2) молекул РНК в пространстве первых трех главных компонент

Успешно выполнен анализ молекул РНК с целью точного определения ОРС. Обучающая выборка ОРС-кандидатов двух типов включала 75 % исходных данных, тестируемая – 25 %. Точность классификации истинных (кодирующих) и псевдо (некодирующих) ОРС-кандидатов – 99,35 %. Оценена информативность признаков фрагментов нуклеотидных последовательностей молекул с использованием критерия на основе индекса Джини [9], встроенного в алгоритм случайного леса. Наиболее информативными признаками являются признаки модели Bao и два варианта вычисления длины ОРС (в количестве нуклеотидов и с использованием логарифмирования). Менее информативными признаками являются частоты различных комбинаций нуклеотидов и корреляционные факторы нуклеотидов. Разработанный классификатор применен для нахождения ОРС 113063 кодирующих молекул РНК (с известными ОРС). Точность нахождения – 98,14 % (точно определены ОРС 110959 молекул).

Выводы. Разработан и успешно проверен вычислительный подход к определению ОРС кодирующих молекул РНК на основе алгоритмов векторизации и случайного леса, обученного на ложных ОРС некодирующих РНК и истинных ОРС кодирующих РНК. Определён набор наиболее информативных признаков фрагментов нуклеотидных последовательностей молекул – это признаки модели Bao и два параметра оценки длины ОРС. Точность определения ОРС в рассмотренных молекулах РНК составляет 98,14 %.

Предложенный вычислительный подход может быть использован в прикладных биомедицинских исследованиях, нацеленных на совершенствование дифференциальной диагностики заболеваний человека генетической природы (включая онкологические заболевания) и для улучшения качества построения прогностических моделей течения подобных заболеваний (включая прогнозирование ответа пациента на лечебную терапию).

1. *Mardis E. R.* // Nat. Protoc. 2017. Vol. 12. P. 213-218.
2. *Reuter J. A., Spacek D. V., Snyder M. P.* // Mol. Cell. 2015. Vol. 58. P. 586–597.
3. *Sayers E. W., Agarwala R., Bolton E. E. et al //* Nucleic Acids Res. 2019. Vol. 47. P. D23–D28.
4. *Kang Y. J., Yang D. C., Kong L. et al //*Nucleic Acids Res. 2017. Vol. 45. P. W12-W16.
5. *Закирова В. Р., Сырокваш Д. А., Гилевский С. В. и др.* // Информатика. 2019. Том. 16, № 2. С. 111–120.
6. *Breiman L.* // Machine Learning. 2001. Vol. 45(1). P. 5–32.
7. *Bao J.* // BMC Bioinformatics. 2014. Vol. 15:321. P. 1–15.
8. *Mao R., Raj Kumar P.K., Guo C. et al* // PLoS One. 2014. Vol. 9(8). P. 1–12.
9. *Яцков Н.Н.* Интеллектуальный анализ данных: пособие. Мн. : БГУ, 2014. 151 с.