АВТОМАТИЧЕСКОЕ ОПРЕДЕЛЕНИЕ ОТКРЫТЫХ РАМОК СЧИТЫВАНИЯ В МОЛЕКУЛАХ РНК ЧЕЛОВЕКА С ИСПОЛЬЗОВАНИЕМ АЛГОРИТМОВ ВЕКТОРИЗАЦИИ И КЛАССИФИКАЦИИ

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@bsu.by, grinev_vv@bsu.by, skakun@bsu.by

Разработан вычислительный подход для автоматического определения открытых рамок считывания в большом наборе молекул РНК на основе алгоритмов векторизации нуклеотидных последовательностей и классификации (случайного леса). Проверка работоспособности алгоритмов выполнена на двух наборах молекул РНК из баз данных NCBI RefSeq и Ensembl. Точность определения открытых рамок считывания составляет 98,14%.

Введение

Современные методы полнотранскриптомного секвенирования [1, 2] позволяют точно установить нуклеотидные последовательности молекул РНК, присутствующих в клетке, а также определить их количественное содержание. При этом кодирующий потенциал таких молекул может быть оценен с помощью алгоритмов определения открытых рамок считывания (ОРС), реализованных в программных пакетах NCBI ORFfinder [3] или CPC2 [4]. Однако данные алгоритмы предназначены для одномолекулярного анализа и не позволяют сделать обоснованный выбор одной из открытых рамок считывания в случае множественности таковых в изучаемой молекуле РНК.

В настоящей работе предложен вычислительный подход для автоматического определения ОРС в большом наборе молекул РНК на основе алгоритмов векторизации нуклеотидных последовательностей и классификации (наиболее вероятной ОРС).

I. Вычислительный подход для определения ОРС

Вычислительный подход включает алгоритмы векторизации [5] и случайного леса [6]. Векторизация нуклеотидных 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) нахождение всевозможных ОРС-кандидатов в молекуле; ii) точное определение ОРС с использованием классификационной модели п. 3.

II. Экспериментальные данные и вычислительный эксперимент

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 произведена на полном наборе кодирующих молекул РНК. Для оценки точности определения ОРС используются координаты ОРС молекул, представленные в базе данных NCBI RefSeq. Для дополнительного подтверждения разработанной классификационной модели рассмотрены 63832 кодирующих молекул РНК с точными координатами ОРС из базы данных Ensembl. Процентная оценка точности определения ОРС молекул производится как отношение числа верно классифицированных ОРС молекул к общему числу рассматриваемых молекул.

III. 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 – 14 часов.

Визуализация результатов векторизации ОРС-последовательностей с использованием метода главных компонент [9] представлена на рис. 1. Два класса ложных и истинных ОРС последовательностей разделяются.

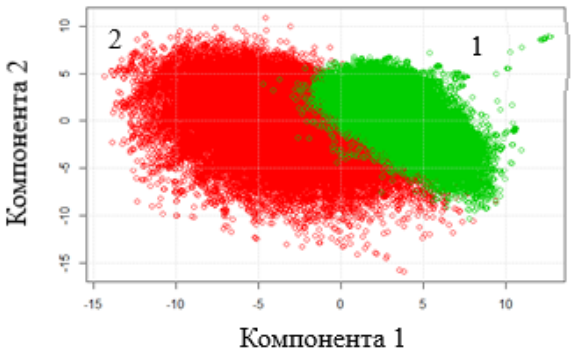

Рис. 1 – Результаты применения метода главных компонент к векторизованному набору данных: ОРС-кандидаты кодирующих (1) и некодирующих (2) молекул РНК в пространстве первых двух главных компонент

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. Для точной классификации фрагментов нуклеотидных последовательностей достаточно использовать 25 наиболее информативных признака. Разработанный классификатор применен для нахождения ОРС 113063 кодирующих молекул РНК (с известными ОРС). Точность нахождения – 98,14% (точно определены ОРС 110959 молекул). Дополнительно проверена работоспособность классификатора на наборе 63832 кодирующих молекул РНК из базы данных Ensembl – точность определения ОРС – 98,14%.

Следует отметить важный сопутствующий вывод по результатам нашей работы, который позволяет предположить о невысокой значимости консенсусных последовательностей Козак [10] при определении координат ОРС и расстояний старт- и стоп-кодонов от 5’ и 3’ начала и конца последовательностей, которые не учтены в разработанной модели векторизации фрагментов ОРС.

Разработан и успешно проверен на большом наборе экспериментальных данных эффективный 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. Отметим, что наши результаты позволяют предположить о невысокой значимости последовательностей Козак при определении координат ОРС и параметров расстояний старт- и стоп-кодонов от 5’ и 3’ начала и конца последовательностей. Точность определения ОРС в рассмотренных молекулах РНК из баз данных NCBI RefSeq и Ensembl 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).

Список литературы

1. Mardis, E. R. DNA sequencing technologies: 2006-2016 / E. R. Mardis // Nat. Protoc. – 2017. – Vol. 12, № 2. – P. 213–218.
2. Reuter, J. A. High-throughput sequencing technologies / J. A. Reuter, D. V. Spacek, M. P. Snyder // Mol. Cell. – 2015. – Vol. 58, № 4. – P. 586–597.
3. Sayers, E. W. Database resources of the National Center for Biotechnology Information / E. W. Sayers, R. Agarwala, E. E. Bolton [et al.] // Nucleic Acids Res. – 2019. – Vol. 47, № D1. – P. D23–D28.
4. Kang, Y. J. CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features / Y. J. Kang, D. C. Yang, L. Kong [et al.] // Nucleic Acids Res. – 2017. – Vol. 45, № W1. – P. W12–W16.
5. Разработка алгоритмов и программных средств классификации кодирующих и некодирующих нуклеотидных последовательностей / В. Р. Закирова [и др.] // Информатика. – 2019. – Т. 16, № 2. – С. 111–120.
6. Breiman, L. Random forest / L. Breiman // Machine Learning. – 2001. – Vol. 45, № 1. – P. 5–32.
7. Bao, J. An improved alignment-free model for DNA sequence similarity metric / J. Bao, R. Yuan, Z. Bao // BMC Bioinformatics. – 2014. – Vol. 15, № 321. – P. 1–15.
8. Comparative analyses between retained introns and constitutively spliced introns in arabidopsos thaliana using random forest and support vector machine / R. Mao [et al.] // PLoS One. – 2014. – Vol. 9, № 8. – P. 1–12.
9. Интеллектуальный анализ данных / Н. Н. Яцков – Минск: БГУ, 2014. – 151 с.
10. Kozak, M. Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes / M. Kozak // Cell. – 1986. – Vol. 44, № 2. – P. 283–292.