КЛАССИФИКАЦИЯ МОЛЕКУЛ РНК

Яцков Н. Н., Климук И. В., 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, ivanklimuk96@gmail.com

Разработаны три модели классификации молекул РНК, на основе: 1) алгоритмов векторизации молекул РНК и классификации случайного леса; 2) определения открытых рамок считывания в молекулах РНК; 3) кодирования единицей («one hot encoding») с последующим использованием одномерной сверточной нейронной сети. Проверка работоспособности разработанных алгоритмов выполнена на примере набора РНК человека из базы данных NCBI RefSeq. Точность классификации молекул с использованием разработанных моделей варьируется от 90,4 до 99,8%.

Введение

Транскрипция генома приводит к образованию матричных РНК, а также разнообразных малых и длинных некодирующих РНК [1]. Длинные некодирующие РНК тесно связаны со многими биологическими процессами, такими как многоуровневая регуляция экспрессии генов. Их структура во многом схожа со структурой матричной РНК, что значительно осложняет задачу точного определения вида молекул, основываясь только на нуклеотидной последовательности. Для решения данной задачи требуется разработка классификационной модели, позволяющей эффективно определять кодирующие и некодирующе молекулы РНК. В работе [2] представлена классификационная модель на основе k-меров и сверточных нейронных сетей, точность классификации которых варьируется от 87,97 до 99,63% для молекул РНК мышей и кур соответственно, для организма же человека точность равна 98,72%. Однако данная модель имеет существенные ограничения, к которым можно отнести ресурс-затратные процедуры определения k-меров, сложность статистического анализа в определении значимости k-меров, определённые требования к вычислительным ресурсам.

Целью данной работы является разработка эффективных математических моделей классификации кодирующих и некодирующих молекул РНК человека, устраняющих недостатки модели [2]. Разработаны три классификационные модели на основе следующих алгоритмов: 1) векторизации молекул РНК и классификации случайного леса; 2) определения открытых рамок считывания (ОРС) молекул РНК; 3) кодирования единицей с последующим использованием одномерной сверточной нейронной сети. Проверка работоспособности разработанных алгоритмов выполнена на примере набора РНК человека из базы данных NCBI RefSeq.

I. Модели классификации

Модель 1. Включает алгоритмы векторизации [3] и случайного леса [4]. Векторизация нуклеотидных последовательностей произведена в 104 признака (частоты моно-, ди- и тринуклеотидов [3], параметры модели Bao [5], корреляционные факторы нуклеотидов [6], длинны последовательностей). Обучение модели производится на выборках кодирующих и некодирующих молекул РНК. Очевидными достоинствами модели являются простота, в том числе алгоритмической и программной реализации, минимальные требования к вычислительным ресурсам, интерпретируемость параметров векторизации.

Модель 2. В основе алгоритма классификации лежит модель 1, обученная на выборках истинных и ложных ОРС, полученных из кодирующих и некодирующих молекул, и алгоритм определение ОРС. Если в молекуле определена ОРС, то молекула считается кодирующей, иначе некодирующей. Преимущества модели: 1) простота программной реализации; 2) умеренные требования к вычислительным ресурсам; 3) возможность прямого использования для решения важной задачи молекулярной биологии – определение ОРС молекул.

Модель 3. Модель представляет двухэтапный алгоритм. На первом этапе осуществляется так называемое кодирование единицей («one hot encoding») молекул РНК [7]. На втором этапе используется одномерная сверточная нейронная сеть [7]. Обучение модели производится на выборках кодирующих и некодирующих молекул РНК. Данная модель фактически является улучшенным вариантом модели [2], однако вместо выполнения процедуры параметризации в векторное пространство k–меров выполняется кодирование единицей, эффективность которого, для решения подобной задачи, подтверждена в [7]. Очевидным недостатком данной модели является требование значительных вычислительных ресурсов.

II. Экспериментальные данные и вычислительный эксперимент

В ходе анализа моделей 1-3 рассмотрены 4235 некодирующих молекул РНК и 5000 случайно отобранных кодирующих молекул РНК

из базы данных NCBI RefSeq. Для модели 2 класс псевдо ОРС-последовательностей содержит 109230 нуклеотидных фрагментов, полученных из 4235 некодирующих молекул РНК. Класс истинных ОРС-последовательностей включает 108654 реальных ОРС молекул РНК. Обучающая выборка моделей 1 и 2 двух типов молекул включала 75% исходных данных, тестируемая – 25%. Обучающая выборка модели 3 – 90% исходных данных, тестируемая – 10%. Для модели 3 выбрана обучающая выборка наибольшего размера с целью увеличения точности классификатора (в задачах классификации генетических данных нейронные сети требуют обучающие выборки больших размеров) [9]. Процентная оценка точности классификации производится как отношение числа верно классифицированных молекул к общему числу.

III. Результаты

Вычислительные алгоритмы моделей 1 и 2 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 – 2-5 минут для модели 1 и 20-30 мин для модели 2.

Алгоритм модели 3 реализован на языке программирования Python с использованием открытых библиотек нейронных сетей глубокого обучения PyTorch и анализа данных scikit-learn. Анализ данных выполнен на 24-ядерной виртуальной машине в облачном сервисе Google Cloud. Время вычислений – 10 часов.

Модель 1. Точность классификации кодирующих и некодирующих РНК – 90,35%. Оценена информативность признаков векторизации молекул с использованием критерия на основе индекса Джини [8], встроенного в алгоритм случайного леса. Наиболее информативными признаками являются признаки модели Bao и частоты появления некоторых тринуклеотидов. Менее информативными признаками являются корреляционные факторы нуклеотидов.

Модель 2. Успешно выполнен анализ молекул РНК. Точность классификации – 96,67%. Точность классификации сопоставима с точностью модели [2], однако скорость вычислений существенно выше, требования к использованию вычислительных ресурсов ниже.

Модель 3. Точность классификации нейронной сети с 9-тью одномерными сверточными слоями и 2-3 эпохами обучения классификации – 99,80%, наивысшая среди рассматриваемых моделей. Однако системные требования существенные – одна эпоха обучения длится около 1 часа. Для конвейерного анализа молекул РНК, данные по которым генерируются с помощью транскриптомного секвенирования, требуются огромные системные ресурсы. Еще одним недостатком модели является известная склонность алгоритмов к переобучению, что может привести к ложной классификации молекул РНК в ходе проведения новых экспериментальных исследований.

IV. Заключение

Разработаны три модели классификации кодирующих и некодирующих молекул РНК, точность которых варьируется от 90,4 до 99,8%. Определён набор наиболее информативных признаков молекул РНК – это признаки модели Bao и набор частот тринуклеотидов.

Модели на основе алгоритмов векторизации молекул и классификации случайного леса практически не уступают в точности опубликованной модели [2], однако существенно превосходят по вычислительной производительности, статистической значимости и интерпретируемости параметров, требуют меньше вычислительных ресурсов.

Классификатор на основе одномерной сверточной нейронной сети является наилучшим, однако требует значительные вычислительные ресурсы (включающие мощные графические карты или многопроцессорные вычислительные серверы (более 24-х ядер)) и имеет склонность к переобучению.

Список литературы

1. Djebali, S. Landscape of transcription in human cells / S. Djebali, C. A. Davis, A. Merkel [et al.] // Nature. – 2012. – Vol. 489, – P. 101–108.
2. Wen, J. Classification model for lncRNA and mRNA based on k-mers and a convolutional neural network / J. Wen, Y. Liu, Y. Shi, H. Huang [et al.] // BMC Bioinformatics. – 2019. – Vol. 20, № 1:469. – P. 1–14.
3. Разработка алгоритмов и программных средств классификации кодирующих и некодирующих нуклеотидных последовательностей / В. Р. Закирова [и др.] // Информатика. – 2019. – Т. 16, № 2. – С. 111–120.
4. Breiman, L. Random forest / L. Breiman // Machine Learning. – 2001. – Vol. 45, № 1. – P. 5–32.
5. 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.
6. 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.
7. Al-Ajlan, A. CNN-MGP: Convolutional Neural Networks for Metagenomics Gene Prediction / A. Al-Ajlan, A. El Allali // Interdiscip Sci. – 2018. – DOI: 10.1007/s12539-018-0313-4.
8. Интеллектуальный анализ данных / Н. Н. Яцков – Минск: БГУ, 2014. – 151 с.
9. Chen, D. Deep learning and alternative learning strategies for retrospective real-world clinical data / D. Chen, S. Liu, P. Kingsbury [et al.] // npj Digital Medicine. – 2019. – Vol. 2, № 43. – P. 1–5.