КЛАССИФИКАЦИЯ МОЛЕКУЛ РНК С ИСПОЛЬЗОВАНИЕМ МОДЕЛЕЙ ВЕКТОРИЗАЦИИ

М. И. Будько

Белорусский государственный университет, г. Минск;
budzko_marie@mail.ru;
науч. рук. – Н. Н. Яцков, канд. физ.-мат. наук, доц.

В работе программно реализованы алгоритмы векторизации и классификации РНК последовательностей, работоспособность которых демонстрируется на примере смоделированных данных. Разработаны: i) имитационная модель генерации данных; ii) модели векторизации последовательностей РНК на основе частот моно-, би- и триграммов нуклеотидов, параметров модели частот и позиций сочетаний нуклеотидов, длин последовательностей, корреляционных факторов нуклеотидов; iii) шесть методов классификации. Результаты работы могут использоваться для исследования экспериментальных данных.

Ключевые слова: нуклеотидная последовательность; векторизация; классификация; кодирующие и некодирующие РНК

ВВЕДЕНИЕ

У «высших» организмов в цепях молекул ДНК и РНК идет чередование интронных и экзонных отрезков [1]. При производстве молекул белков используются лишь интронные отрезки кодирующих молекул РНК. Если идентифицировать кодирующие и некодирующие РНК, то можно значительно сократить размер данных для исследования. Поэтому одна из важных задач биоинформатики – определение кодирующих и некодирующих молекул РНК. Для точной классификации молекул требуется выделить признаки последовательности с использованием алгоритмов векторизации. К основным недостаткам существующих моделей можно отнести: невысокую точность классификации на основе полученных признаков (около 90 %), что существенно меньше, чем при прямом применении нейронных сетей (99 %); ориентированность на обработку определенного типа последовательностей; ограниченно учитывают порядок следования нуклеотидов. Проанализировав наилучшие из существующих модели векторизации, можно разработать улучшенную модель векторизации, устраняющую вышеизложенные недостатки. Цель работы – разработать и исследовать алгоритмы определения кодирующих и некодирующих молекул РНК с использованием моделей векторизации нуклеотидных последовательностей и алгоритмов классификации.

МОДЕЛИРОВАНИЕ ДАННЫХ

Блок-схема алгоритма имитационного моделирования кодирующих и некодирующих молекул РНК представлена на рисунке. За основу взяты две родительские РНК последовательности (блок 1). Длина последовательности *l* генерируется на основе нормального распределения с математическим ожиданием μ и среднеквадратическим отклонением σ.

Моделируемая последовательность разбивается на *m* отрезков длиной в 1 % от заданной длины (длина минимальной интронной последовательности равна 20, берем приблизительно в два раза больше) (блок 2).

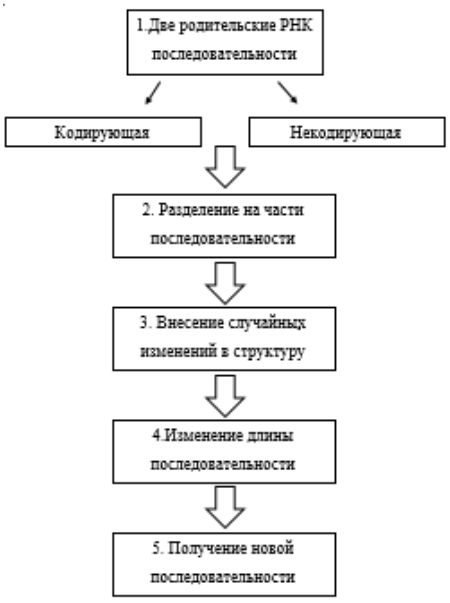

Рис. 1. Блок-схема процесса моделирования данных

В блоке 3 вносятся модификации в последовательности. Выбирается количество изменяемых частей *p* (нормальное распределение). Моделируется число отрезков. Генерируется количество изменений на данных частях последовательности *k* (нормальное распределение). Производится случайная замена (мутация) нуклеотидов: *k* нуклеотидов на каждом из *p* отрезков.

Вносятся изменения в длины результирующих последовательностей (блок 4). Последовательность урезается либо дополняется. В итоге получаем новые последовательности (блок 5).

Произведено моделирование 300 кодирующих и 300 некодирующих РНК последовательностей с перекрытием до 15 % и до 40 %.

ВЕКТОРИЗАЦИЯ ПОСЛЕДОВАТЕЛЬНОСТЕЙ РНК

С помощью процедуры векторизации символьная строка молекулы РНК преобразуется в вектор признаков, что позволяет применять алгоритмы классификации.

Разработаны 3 опорные модели векторизации [2, 3]. Признаки, реализованные в моделях: частоты би- и триграммов нуклеотидов (модель 3); параметры модели частот и позиций сочетаний нуклеотидов (модель 2); длина последовательности (модель 1); корреляционные факторы нуклеотидов (модель 1). Из модели 2 путем нормировки на длину последовательности получена еще одна модель векторизации. Для визуализации данных применен метод главных компонент (МГК). Точность методов классификации оценивается как отношение количества правильно клас-

сифицированных последовательностей РНК к общему количеству объектов из тестовой выборки.

КЛАССИФИКАЦИЯ

В качестве методов классификации выбраны наиболее популярные в литературе методы: *k*-ближайших соседей, AdaBoost, случайного леса, наивная байесовская классификация, опорных векторов, наименьшего расстояния. Методы *k*-ближайших соседей, наименьшего расстояния и наивная байесовская классификация просты в реализации. Методы случайного леса и AdaBoost являются наиболее перспективными алгоритмами классификации. Метод опорных векторов один из наилучших методов для решения задачи бинарной классификации [4].

Выбран язык программирования R, являющийся языком высокого уровня с открытым исходным кодом. Важнейшие достоинства языка R – это открытость и простота изучения, к недостаткам можно отнести низкую производительность сложных алгоритмов [5].

РЕЗУЛЬТАТЫ КЛАССИФИКАЦИИ

Вычислительный эксперимент проводился на двух наборах данных: с перекрытием классов до 15 % и до 40 % (близко к реальным данным). Результаты работы классификаторов отражены в табл. 1 и 2.

Наибольшей точностью обладает метод опорных векторов, близкие показатели имеют AdaBoost и метод случайного леса. Наилучшего значения точности (100 %) при моделировании данных с пересечением до 40 % удалось достичь при использовании модели 2 с нормировкой по длине последовательности, а до 15 % – модели 3 (характерны два четко сформированных класса, с небольшим межклассовым расстоянием).

Таблица 1.

Точность классификации данных с перекрытием до 15 %

	Метод минимального расстояния	Нивная байессовская классификация	Метод *k*-ближайших соседей	Опорных векторов	Случайный лес	AdaBoost
Модель 1	57,8	84,4	89,4	97,8	92,2	94,4
Модель 2	57,8	50,2	88,3	95,6	87,2	88,3
Модель 2, нормиро - ванная по длине	64,4	48,6	100	63,3	64,4	63,3
Модель 3	88,3	89,2	98,3	100	97,8	98,3

Установлено, что для каждой модели подходят свои группы алгоритмов классификации. Например, для модели 3 характерны разделимые классы, между которыми легко провести гиперплоскость, а, следовательно, лучше подходит метод опорных векторов.

Для РНК последовательностей, которые векторизированы с помощью модели 2 (нормированной по длине), лучшие результаты показал метод *k*-ближайших соседей (точность 100 %). Возможно, что объекты данных достаточно близки и внутриклассовое расстояние невелико.

Таблица 2.

Точность классификации данных с перекрытием до 40 %

	Метод минимального расстояния	Нивная байессовская классификация	Метод *k*-ближайших соседей	Опорных векторов	Случайный лес	AdaBoost
Модель 1	55,6	81,6	89,4	95	86,1	88,3
Модель 2	55,6	51,7	88,3	91,7	82,2	78,3
Модель 2, нормиро - ванная по длине	91,6	92,2	100	95	96,1	96,1
Модель 3	62,2	77	98,3	98,8	98,8	97,9

ЗАКЛЮЧЕНИЕ

Разработаны 4 модели векторизации последовательностей нуклеотидов молекул РНК. Предложена имитационная модель генерации кодирующих и некодирующих молекул РНК. Наилучшие результаты получены для модели на основе частот биграммов и триграммов нуклеотидов. Разработаны 6 алгоритмов классификации данных. Максимальная точность 100 % получена для метода опорных векторов. Хорошие результаты показали методы случайного леса и AdaBoost.

Библиографические ссылки

1. *Al-Ajlan A., El Allali* A. Feature selection for gene prediction in metagenomic fragments // BioData mining. 2018. Vol. 11. Art. № 9. DOI: 10.1186/s13040-018-0170-z.
2. *Bao J.* An improved alignment-free model for DNA sequence similarity metric // BMC Bioinformatics / Bao J., Yuan R., Bao Z. // 2014. Vol.15:312. P.1–15.
3. Comparative analyses between retained introns and constitutively spliced introns in arabidopsos thaliana using random forest and support vector machine / Mao R. and others // 2014. Vol. 9, P. 1–12.
4. *Zaki M. J., W. Meira Jr*. Fundamentals of Data Mining Algorithms // Cambridge University Press. 2010. P. 469–546.
5. *Мастицкий С. Э., Шитиков В.К*. Статистический анализ и визуализация данных с помощью R. 2015. Vol. 2. P.31–61.