РАЗРАБОТКА АЛГОРИТМОВ И АВТОМАТИЗИРОВАННЫХ ПРОГРАММНЫХ СРЕДСТВ ДЛЯ КЛАССИФИКАЦИИ КОДИРУЮЩИХ И НЕКОДИРУЮЩИХ НУКЛЕОТИДНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ

Д. А. Сырокваш, Н. Н. Яцков, П. В. Назаров, В. В. Скакун

Белорусский государственный университет
Минск, Беларусь
Люксембургский институт здоровья
Штрассен, Люксембург
e-mail: dasyrokvash@gmail.com

Классификация кодирующих и некодирующих нуклеотидных последовательностей фрагментов молекул ДНК генома человека – наиболее обсуждаемый вопрос в современной биоинформатике. Важной задачей при классификации нуклеотидных последовательностей без выравнивания является выбор модели построения вектора признаков. В ходе работы выявлен ряд статистических закономерностей, которые позволяют различать тип нуклеотидной последовательности, а именно частоты мотивов TA, CG, AAA, TTT, длина последовательности, признаки начала последовательности с мотивами CTG, CTA, GTG или GTA. Точность классификатора на основе метода «случайного леса» с применением данных признаков составила 7,7 %.

Ключевые слова: биоинформатика; ДНК; экзон; интрон; классификация; метод «случайного леса»; R.

DEVELOPMENT OF ALGORITHMS AND AUTOMATED SOFTWARE FOR THE CLASSIFICATION OF THE CODING AND NONCODING NUCLEOTIDE SEQUENCES

D. A. Syrakvash, M. M. Yatskou, P. V. Nazarov, V. V. Skakun

Belarussian State University
Minsk, Belarus
Luxembourg Institute of Health
Strassen, Luxembourg

Classification of coding and non-coding nucleotide DNA sequences of a human genome is an important problem of modern bioinformatics. Choosing an optimal vectorization model is one of most difficult steps of classification of nucleotide sequences without alignment. Statistical estimations of nucleotide sequences have been performed. There are several characteristics of nucleotide sequences that can be used as features for classification: frequencies of TA, CG, AAA, TTT subsequences, length of sequence, flags for sequences starting with subsequences CTG, CTA GTG or GTA.

Coding and non-coding nucleotide sequences classifier based on Random Forest method has been developed. The classifier has error rate 7,7 %.

Keywords: bioinformatics; DNA; exon; intron; classification; Random Forest; R.

Общеизвестным фактом является то, что генетические данные живых мов хранятся в ДНК. Так, в организме человека весь генетический материал хранится в 23 парах хромосом, каждая из которых состоит из огромного числа отдельных стков – генов. В 1977 году в работах Ф. Шарпа [1] и Р. Робертса [2] было показано, что гены высших организмов имеют «прерывистую» структуру: кодирующие отрезки гена перемежаются с некодирующей ДНК, которая не используется при чтении генов. Последнее указало, что возможно сокращение объема исследуемых данных, выделяя из всей структуры гена лишь кодирующие участки. Кодирующие последовательности получили название «экзоны», а некодирующие – «интроны». Важной задачей в ходе классификации нуклеотидных последовательностей является выбор модели ния вектора признаков последовательности и алгоритма. Однако существующие ритмы имеют ряд ограничений [3]: ориентированы на анализ определенного типа последовательностей, слабо учитывают порядок следования нуклеатидов, имеют сокую точность классификации. Решение данной задачи требует разработки и зования усовершенствованных моделей построения вектора признаков и алгоритмов классификации, моделей, автоматизированных программных средств анализа ных данных.

Целью работы является разработка алгоритмов и программного обеспечения для анализа и классификации кодирующих и некодирующих нуклеотидных последовательностей генома человека.

В работе выполнено исследование статистических закономерностей в ДНК/РНК последовательностях организма человека и возможность их использования в качестве признаков для классификации экзонов и интронов. Для разделения набора последовательностей на экзоны и интроны предложена новая модель формирования вектора признаков нуклеотидной последовательности. Признаки новой модели сформированы на основе статистических свойств кодирующих и некодирующих последовательностей. Проведено сравнение ряда моделей векторизации (частотная модель, модель Category-Position-Frequency), модель на основе статистических характеристик) применительно к задаче классификации кодирующих и некодирующих последовательностей.

МОДЕЛИ ПОСТРОЕНИЯ ВЕКТОРА ПРИЗНАКОВ

Основную сложность при решении задачи классификации нуклеотидных довательностей представляет вопрос о сравнении двух последовательностей между собой. Наиболее простым и часто применяемым решением является использование частот встречаемости отдельных подоследовательностей в основной сти [3, 4]. Как правило, в качестве подпоследовательностей выбираются все возможные пары (биграммы) либо тройки (триграммы) нуклеотидов. Полученные 16 либо 64 числовых значения используются в качестве вектора признаков новой модели.

В работах [3] и [4] показано, использование частот встречаемости отдельных символов или небольших групп символов в качестве признаков последовательности нецелесообразно, так как данный подход не учитывает информацию о положении символов в последовательности. В [3] предложена модель CPF (Category-Position-

Frequency), в которой осуществляется формирование из исходной последовательности 3 новых символьных нуклеотидов, 12 новых числовых последовательностей, зования в последовательности локальных частот, вычисление частичных сумм и нововской энтропии последовательностей. Используя полученные значения энтропий как вектор признаков нуклеотидной последовательности авторам, удалось достичь существенного улучшения результатов классификации.

Однако алгоритм формирования вектора признаков модели CPF избыточен, так как правила формирования новых символьных последовательностей из одной последовательности с алфавитом из четырех символов формируют три новых с алфавитом в два символа. В данной работе выполнено исследование возможности понижения размерности модели CPF за счет удаления избыточных компонент.

По результатам исследования статистических характеристик построена новая модель векторизации, признаками которой являются логарифм длины последовательности, частоты биграммов TA и CG, частоты триграммов AAA и TTT, а флаги начала последовательности с триграммов CTG, CTA, GTG и GTA.

СТАТИСТИЧЕСКИЕ СВОЙСТВА НУКЛЕОТИДНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ

В данной работе выполнено исследование статистических свойств набора хромосом организма человека, полученных с электронного ресурса www.ensembl.org. Исходные данные представлены в формате FASTA-файла с набором нуклеотидов и GTF-файла с метаданными. Проведено исследование 22 пар аутосом и X,Y хромосом, общее число обработанных последовательностей составило 2 127 864, из них 1 162 077 экзонов и 965 787 интронов.

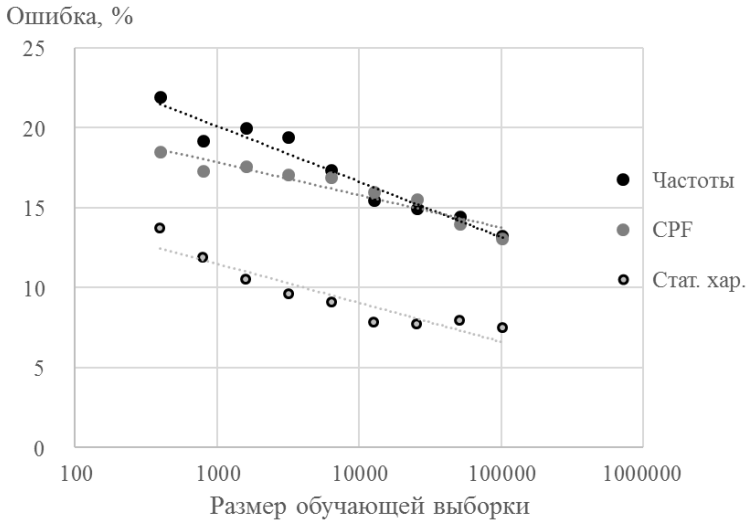
В ходе ранее проведенного исследования длин последовательностей [5] определено, что закон распределения кодирующих и некодирующих нуклеотидных вательностей имеет вид log-нормального распределения. Средняя длина экзона составляет 252 символа, а средняя длина интрона 6540, что говорит о высокой значимости длины последовательности как классифицирующего признака.

Проведено исследование частот встречаемости пар нуклеотидов в экзонах и интронах. В результате данного исследования выделены 5 пар нуклеотидов, для которых наблюдаются значительные отклонения в средних значениях частот между экзонами и интронами: AA, TA, TT, GC, CG. Однако отклонение в частоте мотива GC вызвано наличием длинных CG-последовательностей в экзонах (…CGCGCGCG…) и обусловлено высоким отклонением в частоте мотива CG. Отклонения в частотах мотивов AA и TT вызваны отклонениями в частотах встречаемости более длинных трехсимвольных последовательностей AAA и TTT. Частоты встречаемости биграммов TA и CG, а также триграммов AAA и TTT выбраны в качестве признаков для модели векторизации вследствие того, что именно в этих частотах наблюдается наибольшее отклонение между экзонами и интронами.

Проведено дополнительное исследование начальных символов последовательностей. Обнаружено, что триграммы CTG, CTA, GTG и GTG являются стартовыми подпоследовательностями для 19 % исследованных экзонов и 93 % интронов. Исходя из полученных данных можно сделать вывод, что флаги начала последовательности с триграммов CTG, CTA, GTG и GTA являются значимыми классифицирующими факторами.

Для выявления оптимальной модели формирования вектора признаков было исследованы 4 модели: частотная модель на основе частот биграммов (16 признаков), модель CPF (12 признаков), модель CPF с уменьшенной размерностью (8 признаков), модель на основе статистических характеристик (9 признаков). Для проверки точности классификации сформирована тестовая выборка из 10 000 последовательностей. В качестве параметра оценки точности классификации выбран уровень ошибки классификации как отношения числа неверно классифицированных последовательностей к общему числу последовательностей.

Для получения более точной оценки качества моделей были разработаны классификаторы на основе алгоритма «случайного леса» как наиболее эффективного алгоритма классификации среди существующих на данных момент [6]. В качестве платформы для разработки выбран язык R. Выполнено исследование точности алгоритма в зависимости от объема обучающей выборки. Диапазон изменения объема обучающей выборки – от 100 до 102 400 последовательностей. Зависимость ошибки от объема обучающей выборки приведена на рисунке.



Ошибка классификации в зависимости от размера обучающей выборки для трех моделей векторизации

Следует отметить значительное уменьшение уровня ошибки классификации по сравнению с использованием метода *k*-средних, что подтверждает утверждение о высокой точности метода «случайного леса». Также показано, что наращивание объема обучающей выборки ведет к улучшению точности классификатора.

Наилучший результат классификации показан с применением модели на основе статистических характеристик: уровень ошибок классификации составил 7,7 %, в то время как уровень ошибки классификации при использовании модели CPF составила 13 %. Таким образом, за счет применения новой модели векторизации удалось снизить уровень ошибки в 1,7 раза. Более детальный анализ показал, что данная результат достигнут за счет лучшего распознавания интронов (ошибка 10,3 и 20,5 % соответственно) при примерной равной ошибке распознавания экзонов (5,1 и 6,7 % соответственно).

ЗАКЛЮЧЕНИЕ

В работе представлены результаты исследования статистических закономерностей в ДНК/РНК-последовательностях организма человека и возможность их использования в качестве признаков для классификации экзонов и интронов. Выявлено различие в статистических свойствах кодирующих и некодирующих нуклеотидных последовательностей генома человека. Разработано программное обеспечение для обработки и классификации нуклеотидных последовательностей.

БИБЛИОГРАФИЧЕСКИЕ ССЫЛКИ

1. Sharp P. A., Berk A. J. Sizing and mapping of early adenovirus mRNAs by gel electrophoresis of S1 endonuclease-digested hybrids // Cell 1977. Vol. 12. № 3. P. 721–755.
2. A map of cytoplasmic RNA transcripts from lytic adenovirus type 2, determined by electron microscopy of RNA-DNA hybrids / J. M. Roberts [et al.] // Cell 1977 Vol. 11. № 4. P. 819–855.
3. An improved alignment-free model for DNA sequence similarity metric / J. Bao [et al.] // International Conference on Bioinformatics and Computational Biology BMC Bioinformatics, 2014. Vol. 15 P. 321–336.
4. Li C., Wang J. Relative entropy of dna and its application // Physica A 2005 Vol. 347. № 2. P. 465–471.
5. Сырокваш Д. А., Гунько Е. П. Исследование статистических закономерностей в кодирующих и некодирующих нуклеотидных последовательностях // 72-я научная конференция студентов и аспирантов Белорусского государственного университета, 2015, С. 459–464.
6. Fernandex-Delgado M. Do we Need Hundreds of Classifiers to Solve Real World Classification Problems // J. of Machine Learning, 2014. Vol. 15. P. 3133–3181.