ISSN 1816-0301 (Print)
ISSN 2617-6963 (Online)

БИОИНФОРМАТИКА

BIOINFORMATICS

УДК 57.087.1

Поступила в редакцию 30.08.2018
Received 30.08.2018

Принята к публикации 14.02.2019
Accepted 14.02.2019

Разработка алгоритмов и программных средств классификации кодирующих и некодирующих нуклеотидных последовательностей

В. Р. Закирова[1]✉, Д. А. Сырокваш[1], С. В. Гилевский[1], П. В. Назаров[2], Н. Н. Яцков[1]

[1]*Белорусский государственный университет, Минск, Беларусь*
✉*E-mail: veranika.zakirava@gmail.com*
[2]*Люксембургский институт здоровья, Штрассен, Люксембург*

Аннотация. Проведено исследование кодирующих и некодирующих нуклеотидных последовательностей референсного генома человека. Разработаны семь моделей векторизации нуклеотидных последовательностей на основе частот моно-, би- и триграммов нуклеотидов, параметров модели частот и позиций сочетаний нуклеотидов (category-position-frequency model), длин последовательностей, корреляционных факторов нуклеотидов, статистических признаков кодирующих и некодирующих участков молекул ДНК. Определены наиболее информативные признаки моделей векторизации с использованием алгоритмов автоматического выбора признаков и классификации на основе методов случайного леса и опорных векторов. Установлено различие кодирующих и некодирующих фрагментов нуклеотидных последовательностей. Ошибка классификации последовательностей с использованием метода случайного леса на наборе из 23 наиболее информативных признаков составила 2,93 %.

Ключевые слова: ДНК, экзон, интрон, классификация, метод случайного леса, метод опорных векторов, алгоритмы автоматического отбора информативных признаков, программирование на языке *R*

Для цитирования. Разработка алгоритмов и программных средств классификации кодирующих и некодирующих нуклеотидных последовательностей / В. Р. Закирова [и др.] // Информатика. – 2019. – Т. 16, № 2. – С. 109–118.

Development of algorithms and software for classification of nucleotide sequences

Veranika R. Zakirava[1]✉, Dzmitry A. Syrakvash[1], Stanislau V. Hileuski[1], Petr V. Nazarov[2], Mikalai M. Yatskou[1]

[1]*Belarusian State University, Minsk, Belarus*
✉*E-mail: veranika.zakirava@gmail.com*
[2]*Luxembourg Institute of Health, Strassen, Luxembourg*

Abstract. Coding and non-coding nucleotide sequences of the human reference genome have been investigated. Seven models of vectorization of nucleotide sequences based on mono-, bi-, trigram nucleotide frequencies, parameters of the category-position-frequency model, the lengths of sequences, nucleotide correlation factors,

statistical features of coding and non-coding regions of DNA molecules were developed. The most informative features of vectorization models were determined using feature selection and classification algorithms based on the random forests and support vector machine methods. The difference between coding and non-coding fragments of nucleotide sequences was established. An error of the coding and non-coding sequences classification using the random forests method on a set of the 23 most informative features is 2,93 %.

Keywords: DNA, exon, intron, classification, Random Forests, Support Vector Machine, feature selection, *R* programming

For citation. Zakirava V. R., Syrakvash D. A., Hileuski S. V., Nazarov P. V., Yatskou M. M. Development of algorithms and software for classification of nucleotide sequences. *Informatics*, 2019, vol. 16, no. 2, pp. 109–118 (in Russian).

Введение. Появление новых технологий секвенирования и инструментов для точечного манипулирования структурой ДНК позволяет на генетическом уровне подавлять болезни, повышать устойчивость организмов к неблагоприятным условиям среды и продлевать продолжительность их жизни [1]. В данном контексте определение предназначения генов и их кодирующих и некодирующих участков, экзонов и интронов является одной из первоочередных задач.

Важным этапом обработки нуклеотидных последовательностей является формирование вектора признаков. Существующие модели формирования вектора признаков нуклеотидных последовательностей [2–7] имеют ряд ограничений: они неуниверсальны, в основном предназначены для решения специализированных задач, разработаны для анализа выборок небольшого объема, не включают алгоритмы автоматического выбора признаков [8], что существенно снижает как вычислительную эффективность алгоритмов, так и точность классификации последовательностей вследствие наличия избыточных и неинформативных признаков. Например, исследование способов векторизации нуклеотидных последовательностей для классификации экзонов и интронов представлено в работе [3], однако проведено лишь приближенное сравнение средних значений характеристик классификации на малом объеме данных (менее 10 000 последовательностей) без учета информативности признаков нуклеотидных последовательностей. Перспективным направлением повышения эффективности и точности классификации нуклеотидных последовательностей является отбор их наиболее информативных признаков.

Цель исследования заключается в разработке статистического подхода и программного пакета для классификации кодирующих и некодирующих нуклеотидных последовательностей геномных данных с учетом отбора наиболее информативных признаков нуклеотидных последовательностей. В качестве исходных данных используются опубликованные файлы референсного генома человека [9].

Экспериментальные данные. Рассмотрены нуклеотидные последовательности генома человека [9]. Файлы *gencode.v1.annotation.gtf* и *GRCh38.genome.fa* содержат информацию о биологических последовательностях.

Файл *gencode.v1.annotation.gtf* (*GeneralTransferFormat*) имеет размер 1,13 ГБ и включает:

− имя последовательности;
− источник данных или название программы – генератора данных;
− тип последовательности;
− позицию начала последовательности в *FASTA*-файле;
− позицию конца последовательности в *FASTA*-файле;
− направление прочтения последовательности;
− позицию начала первого кодона;
− дополнительные признаки [9].

Файл *GRCh38.genome.fa* (формат *FASTA*) имеет размер 2,98 ГБ и содержит восстановленный геном человека. Данные представлены парами строк. В первой строке за символом > следует название последовательности. Во второй строке последовательность посимвольно описывается. В файле представлены пять закодированных символов: *A, T, G, C*, соответствующих нуклеотидам аденину, тимину, гуанину, цитозину, а также символ *N*, обозначающий неопределенные нуклеотиды [9].

Исходные данные разделены на 24 файла, характеризующие 22 аутосомы, *X*- и *Y*-хромосомы. Объем файлов данных – 6,13 ГБ; общее количество последовательностей – 2 127 864, из них 1 162 077 экзонов и 965 787 интронов. Ввиду ограничений вычислительных ресурсов анализ данных для каждой из хромосом проводился отдельно.

Рассмотрен набор данных хромосом 1, 4, 7 и 10, включающий более 456 324 последовательностей. Часть данных использовалась в качестве эталонной выборки объемом 1 000 или 100 000 последовательностей, часть являлась тестируемой выборкой объемом 1 000, 10 000 или более последовательностей.

Разработка алгоритмов и программных средств. *Статистический подход для классификации кодирующих и некодирующих нуклеотидных последовательностей*. Разработанный статистический подход для классификации кодирующих и некодирующих последовательностей реализуется в три этапа: предварительная обработка данных, выбор оптимальной модели векторизации, отбор значимых признаков и формирование на их основе модели классификации (рис. 1).

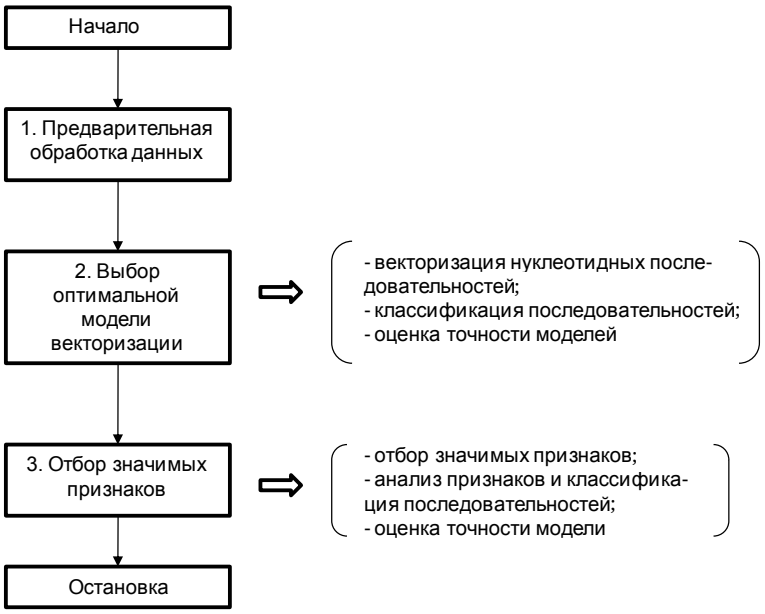

Рис. 1. Общая схема статистического подхода

На этапе предварительной обработки данных осуществляется разбиение исходной последовательности на неперекрывающиеся интронные и экзонные участки, результатом работы является *.csv*-файл со списком обнаруженных последовательностей и указанием их типов.

В ходе второго этапа производится векторизация нуклеотидных последовательностей интронов и экзонов с целью последующего применения алгоритмов классификации. Векторизация подразумевает переход от символьных последовательностей к набору векторов или признаков, характеризующих нуклеотидные последовательности. Рассмотрены следующие модели формирования вектора признаков:

модель 1 – частотная модель, включающая в качестве признаков частоты моно- и биграммов нуклеотидов (20 признаков);

модель 2 – частотная модель с использованием частот триграммов нуклеотидов (64 признака);

модель 3 – модель на основе частот и позиций сочетаний нуклеотидов *CPF* (*category-position-frequency*) [2] (12 признаков);

модель 4 – модификация модели *CPF* с уменьшенным числом компонентов (8 признаков);

модель 5 – модель на основе общих статистических признаков, таких как длина последовательности, частоты нуклеотидов и корреляционные факторы нуклеотидов [10] (13 признаков);

модели 6 и 7 – две модели на основе статистических признаков экзонов и интронов (9 и 12 признаков). Модель 7 включает набор признаков модели 6, а также длину последовательности и флаги начала позиций с биграммов *CT* и *GT*.

Полный набор признаков семи моделей векторизации последовательностей – 110.

Модели 1 и 2 рассматриваются в качестве базовых, наиболее цитируемых в литературе. Модели 3 и 4 являются перспективными, так как используются для решения аналогичных задач в работе [2]. В данных моделях устранен ключевой недостаток моделей 1 и 2, а именно отсутствие в векторе признаков информации о положении и порядке символов в последовательности.

Модели 5–7 на основе статистических признаков четко ориентированы на решение конкретных задач, они популярны ввиду высокой точности получаемых классификаторов.

В качестве общих признаков моделей выбраны [10]:

– длина последовательности;

– частоты нуклеотидов *A*, *T*, *G*, *C*;

– частоты биграммов *AT* и *GC*;

– корреляционные факторы нуклеотидов.

В качестве специальных признаков для решения проблемы классификации экзонов и интронов выбраны [11]:

– логарифм длины последовательности (логарифмическое преобразование позволяет устранить эффект чрезмерного влияния признака с большой вариацией);

– частоты биграммов *TA* и *CG*;

– частоты триграммов *AAA* и *TTT*;

– флаги начала последовательности с триграммов *CTA*, *CTG*, *GTA*, *GTG*.

На втором этапе осуществляется выбор наиболее оптимальных моделей векторизации и классификатора. В качестве алгоритмов классификации рассмотрены наиболее популярные: метод случайного леса [12–14] и метод опорных векторов с радиальной базисной функцией в качестве ядра [15, 16]. Осуществляется грубая классификация нуклеотидных последовательностей с использованием моделей 1–7.

На третьем этапе статистического подхода производится отбор наиболее информативных признаков моделей 1–7 для точной классификации экзонных и интронных последовательностей. В качестве алгоритмов автоматического выбора наиболее информативных признаков экзонов и интронов используются фильтрующие, оберточные и встроенные алгоритмы.

Оценка качества классификации нуклеотидных последовательностей. Оценкой качества классификатора служит уровень допущенных ошибок *ER* (*error rate*):

$$ER = \frac{N_{12} + N_{21}}{N_{11} + N_{12} + N_{21} + N_{22}} \times 100\,\%,$$

где N_{ij} – число последовательностей типа i (кодирующих/некодирующих), распознанных как последовательность типа j (некодирующих/кодирующих). Ошибка классификации оценивалась по тестируемой выборке. Данную оценку целесообразно использовать в ходе анализа выборок, содержащих сопоставимое количество объектов каждого из классов (экзонов и интронов).

Программная реализация алгоритмов анализа данных. В качестве платформы для расчета признаков нуклеотидных последовательностей выбран язык *R* [17], который является языком высокого уровня с открытым исходным кодом для решения статистических задач. Важнейшие достоинства языка *R* – открытость и простота изучения, к недостаткам можно отнести низкую производительность сложных алгоритмов в силу особенностей языка [17]. Для ускорения программных кодов на отдельных этапах анализа данных используется язык программирования *C++*. Ключевым пакетом языка *R* для обработки больших массивов данных является пакет *data.table*, который включает методы быстрой загрузки, формирования выборок и фильтрации данных. Язык *R* обладает довольно ограниченными возможностями при работе с символьными

строками. На языке *C++* реализован алгоритм подсчета количества вхождений подстроки в строку. Для интеграции программных кодов *C++* в язык *R* использовался пакет *Rcpp*. С целью повышения производительности вычислений модели 2 и 3 реализованы на языке *C++*, что дает почти 100-кратное увеличение скорости вычислений в сравнении с *R*-реализациями.

Разработана программа для загрузки и преобразования данных в удобный для дальнейшей работы формат. В качестве параметров принимаются пути к файлам *.gtf* и *.fasta*, название требуемой хромосомы и имя выходного файла. Результатом работы программы является *csv*-файл со списком обнаруженных последовательностей и указанием их типов (экзон или интрон).

Для реализации классификатора на основе метода случайного леса выбран *R*-пакет *randomForest*, содержащий оригинальный алгоритм автора метода Лео Бреймана (Leo Breiman). Особенностью данной реализации является возможность использования алгоритма для решения задач классификации и регрессии, а также гибкой подстройки внутренних параметров алгоритма и оценки значимости входных параметров с помощью индекса Джини (Gini).

Для реализации классификатора на основе нелинейного метода опорных векторов с радиальной базисной функцией в качестве ядра использовался *R*-пакет *e1071*.

Фильтрующие методы. Фильтрующие методы обрабатывают статистические признаки исследуемого набора данных и анализируют каждый признак независимо от остального набора. Основными достоинствами методов являются быстродействие и невысокие требования к производительности вычислительных ресурсов [18].

В качестве программной реализации фильтрующего метода отбора признаков рассмотрен метод одномерных классификаторов *SBF* (*selection by filtering*) *R*-пакета *caret*. Метод строит одномерные линейные классификаторы для каждого из признаков и вычисляет *p*-значение критерия Фишера для оценки значимости признака. Результатом работы метода *SBF* является набор признаков, ранжированный в соответствии со средними *p*-значениями критериев Фишера, полученными в результате *V*-кратной перекрестной проверки. Метод *SBF* возвращает минимальное значение количества признаков, при котором достигается определенное *p*-значение, например, соответствующее заданной точности классификации.

Оберточные методы. Оберточные методы помимо набора тестовых данных требуют информацию об алгоритме классификации. Идея методов заключается в итеративном построении классификаторов на различных подмножествах признаков с использованием результатов классификации в качестве оценки информативности наборов признаков. Методы обладают высокой точностью и избирательностью, позволяют предсказывать оптимальный набор признаков с учетом особенностей классификатора, однако требуют существенных временных затрат. Время их работы нелинейно возрастает в зависимости от количества признаков, что фактически затрудняет применение оберточных методов для анализа данных, характеризующихся большим количеством признаков [18].

В качестве оберточного метода выбран метод рекурсивного удаления признаков *RFE* (*recursive feature elimination*) пакета *caret* [19]. Метод *RFE* строит диаграмму зависимости точности классификатора от количества признаков, позволяя пользователю выбрать необходимый набор признаков, соответствующий заданной точности классификации.

Набор алгоритмов автоматического отбора признаков используется на этапе 3 разработанного статистического подхода. В качестве примера алгоритма из группы встроенных методов рассмотрен метод случайного леса.

Результаты анализа нуклеотидных последовательностей. Вычислительный эксперимент реализуется в три этапа:

1. Предварительная обработка данных. В результате предварительного этапа анализа сформирован *.csv*-файл, список полей которого включает название гена, тип последовательности (экзон или интрон) и символы последовательности.

2. Выбор оптимальной модели векторизации. Вычислены ошибки классификаторов методов случайного леса и опорных векторов при использовании разработанных моделей векторизации (табл. 1). Объемы обучающей и тестируемой выборок составляют 1 000 и 10 000 последовательностей соответственно.

Таблица 1

Уровень ошибки классификации

Модель	Метод случайного леса	Метод опорных векторов
1	18,94	16,32
2	14,60	11,35
3	17,03	15,24
4	17,11	15,29
5	15,55	18,56
6	10,89	9,70
7	8,10	8,38

Оптимизация параметров алгоритмов классификации с помощью пакета *caret* не привела к существенному улучшению результатов.

Можно сделать ряд важных заключений:

1. Наилучшего значения точности (*ER* 8–11 %) удалось достичь при использовании моделей 6 и 7 на основе статистических признаков нуклеотидных последовательностей, в то время как точность моделей 1–5 значительно ниже (*ER* более 11 %).

2. Модель *CPF* (модель 3) действительно содержит избыточные признаки, так как ее точность сопоставима с моделью 4.

3. При использовании модели 5, содержащей длину последовательности в качестве признака, значительно увеличился процент ошибки классификации с применением машины опорных векторов, что обусловлено известной неустойчивостью алгоритма к классификации нестандартизированных данных в условиях высокого шума.

Уровни точности классификации двух методов практически сопоставимы. На третьем этапе анализа данных используется метод случайного леса.

3. Отбор значимых признаков. Исследование значимости отдельных признаков в рамках моделей 1–7 проведено на выборке из 100 000 последовательностей с помощью коэффициента расщепления Джини для классификаторов, построенных с использованием метода случайного леса, показателя *AUC* (*area under curve*, площадь под кривой рабочей характеристики приемника *ROC*) для метода опорных векторов и *p*-значений критериев Фишера для метода *SBF*. В результате сравнительного анализа алгоритмов автоматического отбора признаков исходный набор из 110 признаков моделей 1–7 сокращен до 27 (табл. 2), оценки информативности которых существенно выше, чем у остального набора признаков. Модели 3 и 4 в таблице не представлены, так как они не содержат значимых признаков.

Таблица 2

Наиболее значимые признаки

Мод. 1	Мод. 2	Мод. 5	Мод. 6	Мод. 7
F_{TA}	F_{AAA}	θ_{AT}	*Log*(*Length*)	*Log*(*Length*)
F_{TG}	F_{TAA}	θ_{AG}	*isCTG*	*isCT*
F_{TC}	F_{TAG}	θ_{AC}	*isCTA*	*isGT*
F_{CA}	F_{TTT}	θ_{TG}	*isGTG*	*isCTG*
F_{CG}	F_{TCG}	θ_{TC}	*isGTA*	*isCTA*
–	F_{GTA}	θ_{GC}	–	*isGTG*
–	F_{CGT}	*Length*	–	*isGTA*
–	F_{CCC}	–	–	–

Дальнейшее исследование наборов из 4, 6, 8, 10, 12, 14, 17, 20, 23 и 27 наиболее значимых признаков выполнено на примере обучающей и тестируемой выборок размеров 100 000 и 1 000 соответственно с использованием метода случайного леса и пятикратной перекрестной проверки (рис. 2). Ошибка классификации уменьшается с 6 до 1 % в зависимости от количества при-

знаков. Наименьший уровень ошибки достигается для набора из 23 признаков. Список 23 наиболее информативных признаков представлен в табл. 3.

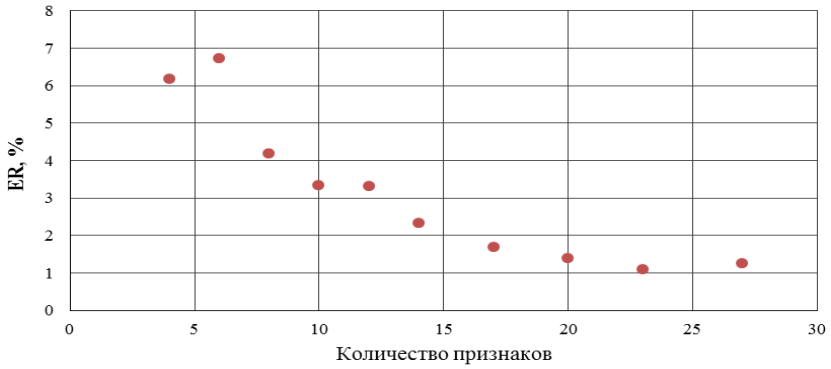

Рис. 2. Зависимость уровня ошибки от количества информативных признаков в результате классификации с использованием метода случайного леса

Признаки длины последовательности *Length* и *Log*(*Length*) наиболее информативные, что очевидно, так как длины интронных и экзонных участков молекул ДНК существенно различаются. В числе наиболее значимых выделены признаки начала последовательности с определенного биграмма (признаки, названия которых начинаются с *is*), корреляционные факторы нуклеотидов, частоты биграммов и триграммов.

Таблица 3

Ранжированный по значениями индекса Джини список признаков

Признак	Ранг *RFE*	Индекс Джини	Признак	Ранг *RFE*	Индекс Джини
Length	1	8574,5	*isGTG*	21	1151,2
Log(*Length*)	2	8274,2	θ_{TG}	5	1141,3
isGT	3	4317,6	F_{CGT}	12	1052,4
isCT	15	3658,2	F_{TG}	14	1048,5
F_{TCG}	10	2082,9	F_{TA}	16	1043,6
F_{TAG}	7	2073,8	F_{TC}	18	1040
isCTG	13	2073	F_{GTA}	9	962,9
θ_{GC}	4	1802,7	θ_{AG}	19	794,9
isGTA	6	1782	F_{CA}	17	794,4
F_{CCC}	8	1761,1	θ_{TC}	23	729,1
F_{CG}	11	1422,6	θ_{AT}	20	652,6
θ_{AC}	22	1340,5	–	–	–

Дополнительно исследованы два набора из 4 и 23 наиболее информативных признаков. Для обучения сформирована выборка размером 100 000 последовательностей, для тестирования сформирован полный набор последовательностей размером 456 324. Ошибки классификации составляют 8,14 и 2,93 % для наборов из 4 и 23 признаков соответственно.

Заключение. В работе предложен статистический подход для классификации кодирующих и некодирующих нуклеотидных последовательностей геномных данных с учетом отбора наиболее информативных признаков нуклеотидных последовательностей. Разработаны и реализованы семь моделей векторизации нуклеотидной последовательности, алгоритмы автоматического выбора признаков, алгоритмы классификации на основе методов случайного леса и опорных векторов.

Проведен анализ экзонных и интронных последовательностей референсного генома человека с использованием разработанных программных средств, по результатам которого выделены 27 информативных признаков моделей векторизации последовательностей.

Выполнено исследование моделей векторизации, алгоритмов классификации и автоматического отбора признаков с целью разделения экзонных и интронных последовательностей на примере анализа аннотированных последовательностей референсного генома человека. Выделены наиболее информативные признаки моделей векторизации последовательностей. Уровень ошибки классификации для наилучшей модели векторизации на основе 23 наиболее значимых признаков составил 2,93 %.

В результате проведенного исследования установлено различие кодирующих и некодирующих фрагментов нуклеотидных последовательностей в референсном геноме человека.

Список использованных источников

1. Edwards, D. J. Beginner's guide to comparative bacterial genome analysis using next-generation sequence data / D. J. Edwards, K. E. Holt // Microbial Informatics and Experimentation. – 2013 – Vol. 3:2. – P. 1–9.

2. 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.

3. Li, C. Relative entropy of DNA and its application / C. Li, J. Wang // Physica A. – 2005. – Vol. 347. – P. 465–471.

4. Numerical characteristics of word frequencies and their application to dissimilarity measure for sequence comparison / Q. Dai [et al.] // J. of Theoretical Biology. – 2011. – Vol. 276. – P. 174–180.

5. Liu, L. Clustering DNA sequences by feature vectors / L. Liu, Y. K. Ho, S. Yau // Mol Phylogenet Evol. – 2006. – Vol. 41. – P. 64–69.

6. Wang, J. Wse, a new sequence distance measure based on word frequencies / J. Wang, X. Zheng // Mathematical Biosciences. – 2008. – Vol. 215. – P. 78–83.

7. Zhao, B. A new distribution vector and its application in genome clustering / B. Zhao, R. L. He, S. T. Yau // Mol Phylogenet Evol. – 2011. – Vol. 59. – P. 438–443.

8. Application of high-dimensional feature selection: evaluation for genomic prediction in man / M. L. Bermingham [et al.] // Scientific Reports. – 2015. – Vol. 5:10312. – P. 1–12.

9. GFF/GTF File Format – Definition and Supported Options [Electronic resource]. – 2014. – Mode of access: www.ensembl.org/info/website/upload/gff.html. – Date of access: 16.10.2014.

10. 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, no. 8. – P. 1–12.

11. Разработка алгоритмов и автоматизированных программных средств для классификации кодирующих и некодирующих нуклеотидных последовательностей / Д. А. Сырокваш [и др.] // Междунар. конгресс по информатике: информационные системы и технологии : материалы конгресса, Минск, 24–27 окт. 2016 г. ; редкол.: С. В. Абламейко [и др.]. – Минск : БГУ, 2016. – С. 189–193.

12. Do we need hundreds of classifiers to solve real world classification problems? / M. Fernández-Delgado [et al.] // J. of Machine Learning Research. – 2014. – Vol. 15. – P. 3133–3181.

13. Liaw, A. Breiman and Custler's Random Forests for Classification and Regression [Electronic resource] / A. Liaw, M. Wiener. – 2016. – Mode of access: http://www.stat.berkley.edu/~breiman/RandomForest/cc_home.htm#workings. – Date of access: 11.02.2016.

14. Breiman, L. Random forest / L. Breiman // Machine Learning. – 2001. – Vol. 45(1). – P. 5–32.

15. Вапник, В. Н. Восстановление зависимостей по эмпирическим данным / В. Н. Вапник. – М. : Наука, 1979. – 448 с.

16. Вьюгин, В. В. Математические основы машинного обучения и прогнозирования / В. В. Вьюгин. – М. : МЦНМО, 2014. – 304 с.

17. Мастицкий, С. Э. Статистический анализ и визуализация данных с помощью *R* [Электронный ресурс] / С. Э. Мастицкий, В. К. Шитиков. – 2014. – Режим доступа: http://r-analytics.blogspot/.com. – Дата доступа: 13.03.2015.

18. Advancing Feature Selection Research – ASU Feature Selection Repository [Electronic resource] / Z. Zhao [et al.]. – 2010. – Mode of access: https://www.researchgate.net/publication/305083748_Advancing_feature_selection_research. – Date of access: 10.04.2019.

19. Kuhn, M. The Caret Package [Electronic resource] / M. Kuhn. – 2017. – Mode of access: https://topepo.github.io/caret. – Date of access: 11.04.2017.

References

1. Edwards D. J., Holt K. E. Beginner's guide to comparative bacterial genome analysis using next-generation sequence data. *Microbial Informatics and Experimentation*, 2013, vol. 3:2, pp. 1–9.
2. Bao J., Yuan R., Bao Z. An improved alignment-free model for DNA sequence similarity metric. *BMC Bioinformatics*, 2014, vol. 15:312, pp. 1–15.
3. Li C., Wang J. Relative entropy of DNA and its application. *Physica A*, 2005, vol. 347, pp. 465–471.
4. Dai Q., Liu X., Yao Y., Zhao F. Numerical characteristics of word frequencies and their application to dissimilarity measure for sequence comparison. *Journal of Theoretical Biology*, 2011, vol. 276, pp. 174–180.
5. Liu L., Ho Y. K., Yau S. Clustering DNA sequences by feature vectors. *Mol Phylogenet Evol*, 2006, vol. 41, pp. 64–69.
6. Wang J., Zheng X. Wse, a new sequence distance measure based on word frequencies. *Mathematical Biosciences*, 2008, vol. 215, pp. 78–83.
7. Zhao B., He R. L., Yau S. T. A new distribution vector and its application in genome clustering. *Mol Phylogenet Evol*, 2011, vol. 59, pp. 438–443.
8. Bermingham M. L., Pong-Wong R., Spiliopoulou A., Hayward C., Rudan I., …, Haley C. S. Application of high-dimensional feature selection: evaluation for genomic prediction in man. *Scientific Reports*, 2015, vol. 5:10312, pp. 1–12.
9. *GFF/GTF File Format – Definition and Supported Options*, 2014. Available at: www.ensembl.org/info/website/upload/gff.html (accessed 16.10.2014).
10. Mao R., Kumar P. K. R., Guo C., Zhang Y., Liang C. Comparative analyses between retained introns and constitutively spliced introns in arabidopsos thaliana using random forest and support vector machine. *PLoS One*, 2014, vol. 9, no. 8, pp. 1–12.
11. Syrakvash D. A., Jackov N. N., Nazarov P. V., Skakun V. V. Razrabotka algoritmov i avtomatizirovannyh programmnyh sredstv dlya klassifikacii kodirujushchih i nekodiruyushchih nukleotidnyh posledovatel'nostey [Development of algorithms and automated software for the classification of coding and non-coding nucleotide sequences]. Mejdunarodnyi congress po informatike: informacionnye sistemy i tehnologii [*International Congress on Informatics: Information Systems and Technologies*]. Minsk, Belorusskij gosudarstvennyj universitet, 2016, pp. 189–193 (in Russian).
12. Fernández-Delgado M., Cernadas E., Barro S., Amorim D. Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, 2014, vol. 15, pp. 3133–3181.
13. Liaw A., Wiener M. *Breiman and Custler's Random Forests for Classification and Regression*, 2016. Available at: http://www.stat.berkley.edu/~breiman/RandomForest/cc_home.htm#workings (accessed 11.02.2016).
14. Breiman L. Random forest. *Machine Learning*, 2001, vol. 45(1), pp. 5–32.
15. Vapnik V. N. Vosstanovlenie zavisimostey po empiricheskim dannym. *Recovering Dependencies from Empirical Data*. Moscow, Nauka, 1979, 448 p. (in Russian).
16. V'ugin V. V. Matematicheskie osnovy mashinnogo obucheniya i prognozirovaniya. *Mathematical Foundations of Machine Learning and Prediction*. Moscow, Moskovskij centr nepreryvnogo matematicheskogo obrazovanija, 2014, 304 p. (in Russian).
17. Mastickiy C. E., Shitikov V. K. Statisticheskiy analiz i vizualizaciya dannyh s pomoshchju *R*. *Statistical Analysis and Data Visualization with R*, 2014. Available at: http://r-analytics.blogspot.com (accessed 13.03.2015) (in Russian).
18. Zhao Z., Sharma S., Morstatter F., Alelyani S. *Advancing Feature Selection Research – ASU Feature Selection Repository*, 2010. Available at: https://www.researchgate.net/publication/305083748_Advancing_feature_selection_research (accessed 10.04.2019).
19. Kuhn M. *The Caret Package*, 2017. Available at: https://topepo.github.io/caret (accessed 11.04.2017).

Информация об авторах

Закирова Вероника Рашидовна, магистрант, кафедра системного анализа и компьютерного моделирования, факультет радиофизики и компьютерных технологий, Белорусский государственный университет, Минск, Беларусь.
E-mail: veranika.zakirava@gmail.com

Information about the authors

Veranika R. Zakirava, Master Student, Department of Systems Analysis and Computer Modelling, Faculty of Radiophysics and Computer Technologies, Belarusian State University, Minsk, Belarus.
E-mail: veranika.zakirava@gmail.com

Сырокваш Дмитрий Алексеевич, магистр, кафедра системного анализа и компьютерного моделирования, факультет радиофизики и компьютерных технологий, Белорусский государственный университет, Минск, Беларусь.
E-mail: dzmitry.syrakvash@gmail.com

Гилевский Станислав Викентьевич, доцент, кандидат технических наук, кафедра системного анализа и компьютерного моделирования, факультет радиофизики и компьютерных технологий, Белорусский государственный университет, Минск, Беларусь.
E-mail: Hileuski@bsu.by

Назаров Петр Владимирович, кандидат физико-математических наук, отдел исследования протеома и генома, Люксембургский институт здоровья, отделение онкологии, Штрассен, Люксембург.
E-mail: petr.nazarov@lih.lu

Яцков Николай Николаевич, доцент, кандидат физико-математических наук, кафедра системного анализа и компьютерного моделирования, факультет радиофизики и компьютерных технологий, Белорусский государственный университет, Минск, Беларусь.
E-mail: yatskou@bsu.by

Dzmitry A. Syrakvash, Master, Department of Systems Analysis and Computer Modelling, Faculty of Radiophysics and Computer Technologies, Belarusian State University, Minsk, Belarus.
E-mail: dzmitry.syrakvash@gmail.com

Stanislau V. Hileuski, Associate Professor, Cand. Sci. (Eng.), Department of Systems Analysis and Computer Modelling, Faculty of Radiophysics and Computer Technologies, Belarusian State University, Minsk, Belarus.
E-mail: Hileuski@bsu.by

Petr V. Nazarov, Cand. Sci. (Phys.-Math.), Proteome and Genome Research Unit, Luxembourg Institute of Health, Department of Oncology, Strassen, Luxembourg.
E-mail: petr.nazarov@lih.lu

Mikalai M. Yatskou, Associate Professor, Cand. Sci. (Phys.-Math.), Department of Systems Analysis and Computer Modelling, Faculty of Radiophysics and Computer Technologies, Belarusian State University, Minsk, Belarus.
E-mail: yatskou@bsu.by