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

БИОИНФОРМАТИКА

BIOINFORMATICS

УДК 51-76[577.21+004.4]; 577.21:519.1

Поступила в редакцию 13.06.2019
Received 13.06.2019

Принята к публикации 31.07.2019
Accepted 31.07.2019

Вычислительный подход и программный пакет RNAexploreR для группировки молекул РНК генов человека по их экзонным признакам

Н. Н. Яцков✉, В. В. Скакун, В. В. Гринев

Белорусский государственный университет, Минск, Беларусь
✉E-mail: yatskou@bsu.by

Аннотация. Изучение правил комбинаторики экзонов генов человека во время сплайсинга представляет огромный интерес для диагностики и лечения раковых заболеваний. Определенная часть исследований направлена на разработку надежных моделей предсказания глобальной комбинаторики экзонов при образовании зрелой РНК. Первоочередной задачей является разработка стандартов или единых систематизированных статистических подходов к анализу и интерпретации возможных экзонных последовательностей генов.

Предложен вычислительный подход к группировке событий альтернативного сплайсинга в первичных матричных РНК генов человека с целью определения генной принадлежности или класса молекул, методика которого состоит в снижении размерности пространства экзонных признаков и объединении близко расположенных экзонов в ограниченное число классов, замене экзонных путей генерации РНК на последовательности соответствующих меток классов экзонов, вычислении расстояний между транскриптами РНК по некоторой мере сходства, объединении близкорасположенных объектов РНК в кластеры. Проверка работоспособности разработанных алгоритмов выполнена на примере наборов молекул РНК отобранных негомологичных генов человека и гибридного онкогена RUNX1/RUNX1T1 человека. Средняя точность отнесения транскрипта к заданному гену составила 99,5 % для рассмотренных негомологичных пар генов.

Разработаны программный пакет и веб-приложение RNAexploreR, интегрирующие реализованные алгоритмы анализа альтернативного сплайсинга РНК-продуктов генов человека. Предложенные алгоритмы и программное обеспечение могут быть использованы для изучения организации и функционирования как аберрантных, так и нормальных генов человека.

Ключевые слова: гены человека, гибридный онкоген RUNX1/RUNX1T1, альтернативный сплайсинг, признаки экзонов, интеллектуальный анализ данных, метод главных компонент, кластерный анализ

Благодарность. Работа выполнена в рамках государственной программы научных исследований «Конвергенция-2020» Республики Беларусь (грант № 3.08.3, номер госрегистрации 20162176).

Для цитирования. Яцков, Н. Н. Вычислительный подход и программный пакет RNAexploreR для группировки молекул РНК генов человека по их экзонным признакам / Н. Н. Яцков, В. В. Скакун, В. В. Гринев // Информатика. – 2019. – Т. 16, № 4. – С. 7–24.

A computational approach and software package RNAexploreR for grouping RNA molecules of human genes by exon features

Mikalai M. Yatskou✉, Victor V. Skakun, Vasily V. Grinev

Belarusian State University, Minsk, Belarus
✉*E-mail: yatskou@bsu.by*

Abstract. The study on the exon combinatoric rules of human genes during the process of splicing is of great interest for the diagnosis and treatment of cancer. A certain part of the research is aimed at developing reliable prediction models for global exon combinatorics during the formation of mature RNA. The primary task is to develop standards or uniform systematic statistical approaches to the analysis and interpretation of possible exon sequences of genes.

A computational approach is proposed to group alternative splicing events in primary messenger RNA of human genes with the aim of determining the gene correspondence or molecule class. The method consists of reducing the dimension of the exon feature space and combining closely located exons into a limited number of classes, replacing the exon pathways of RNA generation with sequences of corresponding exon class labels, calculating the distances between RNA transcripts by some measure of similarity, and associating closely spaced RNA objects into clusters. The performance evaluation of developed algorithms has been done using the examples of RNA molecules of selected nonhomologous human genes and human hybrid oncogene RUNX1/RUNX1T1. The mean accuracy of the assignment of the transcript to given gene is about 99,5 % for the considered nonhomologous pairs of genes.

A software package and web application RNAexploreR, integrating the implemented algorithms for the analysis of alternative splicing of human gene RNA products, have been developed. The proposed algorithms and software can be used to study the organization and functioning of both aberrant and normal human genes.

Keywords: human genes, hybrid oncogene RUNX1/RUNX1T1, alternative splicing, exon features, data mining, principal component analysis, cluster analysis

Acknowledgements. This work was carried out in the framework of the state program of scientific research "Convergence 2020" of the Republic of Belarus (grant no. 3.08.3, number state registration 20162176).

For citation. Yatskou M. M., Skakun V. V., Grinev V. V. A computational approach and software package RNAexploreR for grouping RNA molecules of human genes by exon features. *Informatics*, 2019, vol. 16, no. 4, pp. 7–24 (in Russian).

Введение. Сплайсинг РНК является фундаментальным процессом, протекающим во всех без исключения клетках эукариот и приводящим к образованию из молекул-предшественников зрелых функционально активных РНК-продуктов [1]. Ключевое событие сплайсинга – это распознавание компонентами сплайсингосомы границ интрона (донорного и акцепторного сайтов сплайсинга) с последующим его удалением из молекулы-предшественника и объединением двух экзонов, ранее разделенных таким интроном, с помощью новой фосфодиэфирной связи [2, 3]. Как правило, в теле гена любой из его интронов фланкирован несколькими альтернативными сайтами сплайсинга разной силы. Кроме того, в непосредственной близости от сайтов сплайсинга как в самом интроне, так и в примыкающих к нему экзонах располагается множество сайтов связывания разнонаправленно действующих белков – регуляторов сплайсинга (как энхансеров, так и сайленсеров), причем разной аффинности [4]. Эти белки во многом определяют, будет ли включен тот или иной экзон в состав зрелой РНК или же он будет исключен вместе с примыкающими к нему интронами. В итоге любой из мультиэкзонных генов потенциально мог бы продуцировать огромное разнообразие вариантов зрелых РНК в зависимости от того, по каким именно сайтам сплайсинга шло бы «монтирование» конечной молекулы РНК [5, 6], но в действительности этого не происходит и каждый ген порождает ограниченный набор зрелых РНК-продуктов.

Правила и механизмы сплайсинга, существенно ограничивающие разнообразие зрелых РНК-продуктов гена, до конца не установлены. Наибольший прогресс здесь достигнут при идентификации детерминант и построении предсказательной модели, описывающей включение

или исключение одного кассетного экзона, фланкированного двумя конститутивными экзонами [7]. Однако даже в этом сравнительно простом случае точное предсказание включения или исключения кассетного экзона невозможно без такого комплексного признака, как тканевая специфичность. Надежной модели, которая позволяла бы только на основании сиквенс-признаков молекулы-предшественника предсказывать как локальную (лишь несколько соседних экзонов), так и, самое главное, глобальную (все экзоны конечного варианта РНК) комбинаторику экзонов при образовании зрелой РНК, на сегодняшний день не разработано.

В связи с этим была предпринята попытка разработать такой алгоритм (или методику вычислительного подхода) интеллектуального анализа данных, который позволял бы, опираясь на множество признаков экзонов, а также фланкирующих их интронных последовательностей, надежно разделять или группировать молекулы РНК, имеющие разные структурную организацию и генную принадлежность. Разработанный алгоритм был успешно апробирован на наборе экспериментально подтвержденных и теоретически предсказанных (на основе графа сплайсинга) молекул РНК гибридного онкогена RUNX1/RUNX1T1 человека, а также аннотированных РНК-транскриптов случайно выбранных нормальных генов человека. Авторы полагают, что разработанная алгоритмическая методика является первым существенным шагом на пути расшифровки закономерностей глобальной комбинаторики экзонов, происходящей при образовании зрелых молекул РНК генов человека.

Материалы и методы. В исследование был включен гибридный онкоген RUNX1/RUNX1T1 человека, который является продуктом реципрокной негомологичной транслокации t(8;21)(q22;q22) и появление которого в гемопоэтических стволовых клетках приводит к их перерождению в лейкозные [8]. Кроме того, для исследования были отобраны 14 нормальных (или модельных) генов человека из базы данных Ensembl (выпуск 85, основанный на эталонной сборке GRCh38.p7 генома человека) [9]. В список таких генов вошли семь кодирующих белки генов (ADGRG1, C2orf92, NDRG2, SORBS2, ST5, TCF4 и TEX41) и столько же генов некодирующих РНК (AL157392.3, MIR7515HG, MUC20-OT1, MYO3B-AS1, SATB1-AS1, SOX2-OT и TTN-AS1).

Список экспериментально установленных полноразмерных транскриптов гибридного онкогена RUNX1/RUNX1T1 человека включал 135 изоформ РНК, описанных ранее в работах [5, 6]. Теоретически возможные транскрипты для этого онкогена были рассчитаны путем полного обхода направленного ациклического графа сплайсинга, реконструированного на основе экзон-экзонных стыков, которые были обнаружены в реальных транскриптах гибридного онкогена RUNX1/RUNX1T1. Полный список теоретически возможных транскриптов включал 43 353 изоформы [6]. Кроме того, набор транскриптов включал 892 изоформы РНК отобранных ранее модельных генов человека, аннотированных в базе данных Ensembl.

Все разнообразие транскриптов гибридного онкогена RUNX1/RUNX1T1, отмеченное выше, основано на использовании 99 уникальных вариантов экзонов этого онкогена. Аналогичным образом все разнообразие транскриптов модельных генов порождено комбинаторикой 1762 уникальных экзонов этих генов.

Экзонные признаки. Экзоны, вовлеченные в исследование, были описаны с помощью 1438 численных признаков. Эти признаки ассоциированы со сплайсингом и описывают особенности нуклеотидной последовательности самого экзона или фланкирующих его интронов (такие, как сила сайтов сплайсинга, частота или аффинность сайтов связывания для белков – регуляторов сплайсинга, устойчивость вторичной структуры или эволюционный консерватизм последовательности и т. д.). Подробное описание использованных признаков дано в работе [6].

Методика вычислительного подхода к группировке событий альтернативного сплайсинга. Решение задачи анализа и моделирования транскрипционных данных реализуется с помощью алгоритмов и программных средств интеллектуального анализа данных (data mining) [10–12]. Прямым способом разделения или группировки теоретических молекул РНК является их векторизация путем составления набора данных, в котором для каждой молекулы РНК экзоны заменяются векторами их признаков в случае наличия соответственных экзонов и векторами нулей в случае отсутствия указанных экзонов. Затем решается задача классификации и предсказания типов молекул РНК. Основной недостаток данного подхода заключается в со-

ставлении огромного набора данных, содержащего более 40 000 и 100 000 строк и колонок соответственно (например, для гибридного онкогена RUNX1/RUNX1T1 молекул РНК – 43 353 (число строк), экзонов – 99 и признаков экзонов – 1438 (число столбцов 142 362)), что затрудняет применение вычислительных алгоритмов анализа данных. Более усовершенствованным подходом является предсказание молекул РНК с использованием алгоритмов поиска ассоциативных правил [5]. С помощью данных алгоритмов путем анализа структуры экспериментально обнаруженных транскриптов генов формулируются ассоциативные правила, описывающие наиболее значимые и устойчивые комбинации экзонов. Найденные правила последовательно применяются к списку теоретически предсказанных изоформ РНК, тем самым оставляя в нем только те варианты, которые содержат комбинации экзонов, соответствующие правилам. Однако данный подход не учитывает свойства экзонов, входящих в последовательности молекул РНК, что делает результат анализа слишком грубым, определяющим в основном короткие транскрипты или незначительно отличающиеся (в пределах одного-двух экзонов) от наиболее распространенных вариантов экспериментально подтвержденных РНК. Развитием идеи поиска ассоциативных правил является байесовский подход [10, 13], основанный на расчете вероятностей появления молекул РНК исходя из условных вероятностей появления входящих в них экзонов, которые оцениваются по частотам появления экзонов в выборке экспериментально подтвержденных молекул РНК. К недостаткам подхода можно отнести нахождение коротких молекул, состоящих из наиболее вероятных экзонов, и неучет свойств признаков экзонов, что при грубой оценке условных вероятностей появления экзонов весьма критично и может существенно понизить точность определения молекул РНК. Высокий потенциал имеют подходы, основанные на векторизации нуклеотидных последовательностей молекул РНК и применении нейронных сетей глубокого обучения [14–17]. Однако данные подходы весьма затратны в плане потребления вычислительных ресурсов, так как используют ресурсоемкие вычислительные алгоритмы (специальные алгоритмы векторизации нуклеотидных последовательностей, алгоритмы автоматического отбора наиболее информативных признаков (features selection), сверточные нейронные сети), что существенно ограничивает их широкую применимость. Очевидно, что наиболее перспективным способом решения задачи появления молекул РНК определенного класса является подход, позволяющий выполнить прямое сравнение молекул РНК с учетом свойств входящих в них экзонов. В работе предлагается методика для подобного подхода, включающая набор алгоритмов интеллектуального анализа данных, согласно которой молекулы РНК могут сравниваться напрямую без использования грубых или точных, но ресурсоемких вычислительных алгоритмов.

Идея вычислительного подхода состоит в снижении размерности пространства экзонных признаков и объединении близко расположенных экзонов в ограниченное число классов, замене экзонных путей генерации РНК на последовательности соответствующих меток классов экзонов, вычислении расстояний между транскриптами РНК по некоторой мере сходства, объединении близко расположенных объектов РНК в кластеры. Основные этапы подхода с учетом выбранных наиболее оптимальных алгоритмов интеллектуального анализа данных представлены на рис. 1. Рассмотрим их более подробно.

Этап 1. Снижение размерности пространства признаков экзонов. Очевидно, что бóльшая часть признаков, описывающих нуклеотидные последовательности экзонов генов, малоинформативная. Учет группы малоинформативных признаков приводит к затруднению анализа данных, а именно к их зашумлению, увеличению объема данных, искажению достоверной информации о кластерах схожих экзонов, снижению точности классификации данных. Для улучшения качества разбиения экзонов на кластеры требуется проведение этапа анализа данных, включающего переход от неинформативных атрибутов экзонов к информативным в смысле их разделения на кластеры. Для выполнения данного преобразования требуется применение алгоритмов снижения размерности данных [18]. Снижение размерности данных является наиболее эффективным подходом для удаления шумовых и неинформативных атрибутов объектов. Методы снижения размерности данных включают две группы алгоритмов: на основе построения (не)линейных комбинаций признаков (feature extraction) и на основе выделения наиболее информативных исходных признаков (feature selection). К алгоритмам первой группы

относятся методы главных и независимых компонент, дискриминантный и факторный анализ [18–21]. Идея алгоритмов данной группы состоит в переходе в пространство низкой размерности (новых признаков) без потери сущности информации. Идея алгоритмов второй группы состоит в выделении небольшой группы исходных наиболее информативных признаков объектов, минимизирующих шум и избыточность в данных и максимизирующих их информативность в смысле разделения на кластеры [22, 23]. Часто применение алгоритмов второй группы является предпочтительным, так как не приводит к изменению исходных данных, в то время как в пространстве новых признаков взаимное расположение кластеров данных может измениться, что приведет к неверной биофизической интерпретации исследуемых процессов. Ввиду того что в данной работе в первую очередь требуется сокращение пространства признаков экзонов и увеличение точности их последующей классификации, принято решение об использовании для снижения размерности данных метода линейного преобразования признаков, а именно метода главных компонент. Метод главных компонент служит базовым алгоритмом сжатия данных. Он обладает высоким быстродействием, что обеспечивает проведение вычислительного эксперимента при малых вычислительных затратах, и является относительно простым для программной реализации.

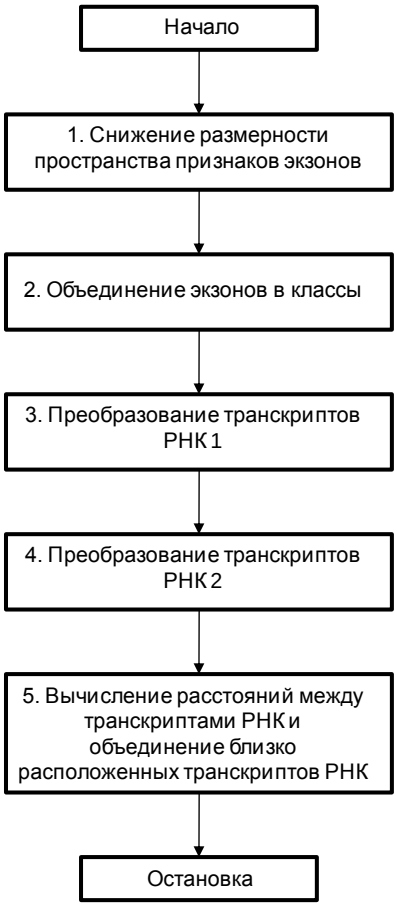


Рис. 1. Блок-схема вычислительного подхода для группировки событий альтернативного сплайсинга в первичных матричных РНК генов человека

Этап 2. Объединение экзонов в классы. Объединение экзонов представляет собой разбиение множества элементов на группы без обучения и является задачей кластерного анализа. Кластерный анализ не требует априорной информации о метках классов данных и позволяет разделить множество исследуемых объектов на группы похожих объектов – кластеры. По способам кластеризации методы кластерного анализа можно условно разделить на две большие группы: иерархические и неиерархические методы [24, 25]. Каждая из групп методов включает множество подходов и алгоритмов. Для кластерного анализа небольших наборов данных экзонов

(до 5000 объектов) предпочтительнее использовать медленные и более точные иерархические методы, для анализа больших наборов данных (более 5000 объектов) – быстрые и менее точные методы неиерархического анализа. В работе применяется иерархическая кластеризация экзонов генов на основе отобранного набора новых признаков (главных компонент). Выполняется разбиение экзонов на кластеры и присвоение каждому кластеру уникального индекса.

Этап 3. Преобразование транскриптов РНК 1. Осуществляется преобразование транскриптов первой из сравниваемых молекул РНК. Для уменьшения вычислительной сложности решения задачи предсказания экзонных последовательностей РНК используется прием разделения пространства экзонов гена на непересекающиеся области, при этом каждой области ставится в соответствие символ или метка класса (индекс). Конечное множество символов образует алфавит исследуемых последовательностей. Выполняется преобразование символов последовательностей экспериментально подтвержденных транскриптов РНК (от имен экзонов) в метки или индексы кластеров, в которых расположены соответствующие экзоны. На данном этапе производится удаление транскриптов-дубликатов. В случае необходимости анализ удаленных транскритпов РНК осуществляется после выполнения завершающего этапа подхода.

Этап 4. Преобразование транскриптов РНК 2. Выполняется преобразование транскриптов второй молекулы РНК в метки кластеров экзонов и удаление транскриптов-дубликатов. Исследование удаленных транскритпов РНК производится после проведения завершающего этапа подхода.

Этап 5. Вычисление расстояний между транскриптами РНК и объединение близко расположенных транскриптов РНК. Суть решаемой задачи состоит в формализации различия между экзонными последовательностями транскриптов РНК и введении отношения расстояния между ними для последующего определения их подобия. При этом принимается ряд допущений: экзоны со схожими характеристиками считаются однотипными, а экзонные последовательности РНК, расположенные на малых расстояниях, – функционально близкими и соответствующими определенным биологическим процессам. Функцией расстояния может являться некоторый алгоритм для вычисления расстояния между символьными строками. Объединение близко расположенных транскриптов РНК производится с помощью методов кластерного анализа. Например, теоретически предсказанные транскипты РНК некоторого гена, попадающие в кластеры экспериментально подтвержденных транскриптов РНК и определенные по некоторому порогу меры расстояний сходства, считаются наиболее вероятными вариантами альтернативного сплайсинга для указанного гена. Точная мера близости предсказанного транскрипта к заданному экспериментально подтвержденному транскрипту оценивается с помощью заданной функции расстояния, или меры сходства [26, 27].

Следует отметить, что алгоритмы кластерного анализа применяются на втором и пятом этапах разработанного подхода: на втором этапе предпочтительнее использовать иерархические алгоритмы кластерного анализа, так как число кластеризуемых объектов невелико (100–200); на пятом этапе – алгоритмы неиерархического кластерного анализа, так как число объектов может быть велико. Однако в неиерархических итерационных алгоритмах требуется многократное вычисление расстояний между объектами данных на каждой итерации, что существенно замедляет анализ данных в случае расчета мер сходства символьных последовательностей. В иерархических алгоритмах достаточно единожды вычислить матрицу расстояний между объектами, которая затем используется в вычислительных процедурах алгоритмов. Поэтому для кластерного анализа транскриптов РНК также выбраны иерархические алгоритмы.

Описание вычислительного эксперимента. Для проверки работоспособности разработанных алгоритмов подхода рассмотрено предсказание событий сплайсинга на примерах экспериментально подтвержденных транскриптов различных пар 14 модельных генов. Случайным образом составлены 10 пар модельных генов (табл. 1). В случае успешной работы вычислительного подхода экспериментально подтвержденные транскрипты РНК различных генов должны быть предсказаны с высокой точностью, т. е. должны сформироваться различные кластеры данных. Анализ составленных пар генов выполнен в соответствии с этапами разработанного подхода.

Таблица 1

Пары сравниваемых 14 модельных генов

Номер пары	Ген 1	Ген 2	Количество экспериментально подтвержденных транскриптов (ген 1 / ген 2)	Количество экзонов, формирующих экспериментально подтвержденные транскрипты (ген 1 / ген 2)
1	ENSG00000228956	ENSG00000205336	65 / 72	118 / 156
2	ENSG00000154556	ENSG00000166444	57 / 50	130 / 121
3	ENSG00000165795	ENSG00000196628	55 / 61	106 / 140
4	ENSG00000226674	ENSG00000228486	78 / 55	126 / 93
5	ENSG00000231898	ENSG00000236172	67 / 53	127 / 107
6	ENSG00000237298	ENSG00000239665	58 / 50	127 / 93
7	ENSG00000242086	ENSG00000242808	121 / 50	221 / 97
8	ENSG00000228956	ENSG00000165795	65 / 55	118 / 106
9	ENSG00000205336	ENSG00000226674	72 / 78	156 / 126
10	ENSG00000154556	ENSG00000231898	57 / 67	130 / 127

Анализ полного набора признаков экзонов на этапе 1 выполнен с использованием метода главных компонент и процедуры стандартизации данных. Процедура стандартизации позволяет устранить чрезмерный эффект влияния признаков с наибольшим разбросом. Реализован отбор первых главных компонент, объясняющих требуемую долю вариации в данных (например, 95 или 99 %).

Для кластерного анализа экзонов генов выбран иерархический агломеративный кластерный анализ. В качестве расстояний между объектами рассмотрены: евклидово (euclidean), Чебышева (maximum), городских кварталов (manhattan) и Минковского (minkowski, p = 4); в качестве методов связывания – Уорда (ward), ближайшего (single) и дальнего (complete) соседей, средней связи (average), Маккуити (mcquitty) и центроидное (centroid) [24, 25]. Оптимальное число кластеров экзонов может быть определено в результате дополнительного исследования количества кластеров экзонов, варьируемого от некоторого минимального (например, пять) до максимального значения числа экзонов. Нижняя граница для числа кластеров определяется условием установленной точности классификации, верхняя – вычислительными возможностями исследователя. В данной работе разбиение экзонов производится на 26 кластеров. Это количество является оптимальным в первом приближении для проверки работоспособности предлагаемого подхода и соответствует количеству символов латинского алфавита, что упрощает алфавитное представление экзонов последовательностей РНК в смысле общепринятых лексических ассоциаций.

Для определения наиболее эффективного метрического расстояния и способа связывания объектов в алгоритмах иерархического кластерного анализа рассмотрены кофенетический корреляционный коэффициент κ [28] и обратная величина расчетной статистики критерия χ^2 [29]. Два наугад выбранных объекта *l* и *m* на иерархическом дереве связаны между собой так называемым кофенетическим расстоянием ρ_{lm}. Величина этого расстояния определяется расстоянием между двумя кластерами, в которых находятся данные объекты. Мерой линейной связи между кофенетическими ρ и метрическими расстояниями *r* между объектами данных является кофенетический корреляционный коэффициент κ:

$$\kappa = \frac{\mathrm{cov}_{\rho r}}{\sigma_\rho \times \sigma_r}, \qquad (1)$$

где $\mathrm{cov}_{\rho r}$ – ковариация между кофенетическими ρ и метрическими расстояниями *r*; σ_ρ и σ_r – среднеквадратические отклонения для оценок кофенетических ρ и метрических расстояний *r*.

Построение иерархического дерева считается успешным, если кофенетический корреляционный коэффициент близок к единице. К недостатку критерия оценки качества на основе кофенетического корреляционного коэффициента следует отнести неспособность учета равномерности заполнения кластеров данных. Равномерность заполнения кластеров данных снижает эффект множественного появления транскриптов-дубликатов (ввиду наличия большого количества экзонов с определенной меткой класса), что существенно повышает точность последующей классификации транскриптов РНК. Для оценки равномерности заполнения кластеров данных используется величина

$$\chi^2 = \sum_{j=1}^{J} \frac{\left(v_j - n \times p_j\right)^2}{n \times p_j}, \tag{2}$$

где v_j – число экзонов в j-м кластере; $p_j = 1/J$; J – число кластеров; n – число экзонов.

Для оценки качества иерархического кластерного анализа рассмотрен интегральный коэффициент качества кластерного анализа Q, представляющий собой среднее арифметическое кофенетического корреляционного коэффициента и нормированной обратной величины χ^2:

$$Q = \frac{1}{2} \times \left(\kappa_{ij} + k \times \frac{1}{\chi_{ij}^2} \right), \tag{3}$$

где κ_{ij}, χ_{ij}^2 – кофенетический корреляционный коэффициент и величина критерия χ^2 для вычисления расстояния по некоторой мере сходства i и способу связывания j. Нормировочный коэффициент k выбирается из условия приведения к единой шкале двух критериев. В данной работе $k = 100$. Примеры оценки качества иерархического кластерного анализа показаны на рис. 2.

В качестве мер сравнения последовательностей транскриптов РНК генов на этапе 4 рассмотрены расстояния Джаро – Винклера (jw), оптимального совпадения строк (osa), Левенштейна (lv), Дамерау – Левенштейна (dl), наибольшей общей подстроки (lcs), q-грамм (qgram), косинусное (cosine), Жаккара (jaccard) и расстояние кодирования в фонетический код (soundex) [27, 30–32].

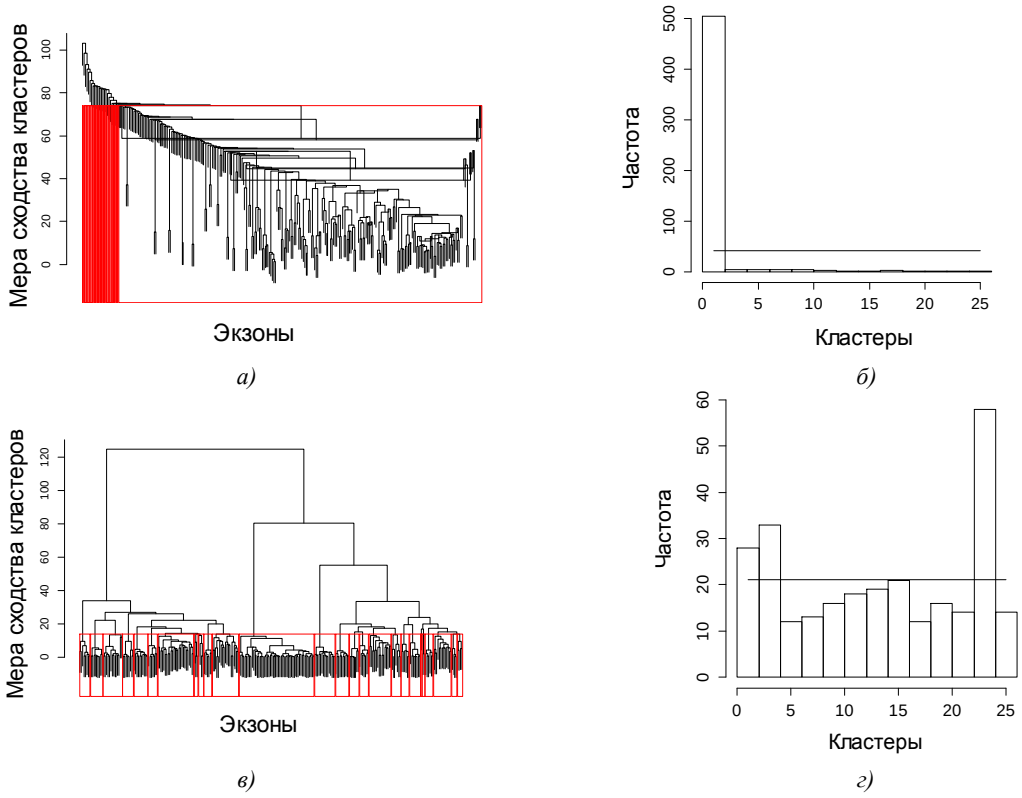
Для реализации процедуры кластерного анализа транскриптов молекул РНК различных генов, представленных матрицей вычисленных расстояний между транскриптами, выбран иерархический агломеративный кластерный анализ. В качестве связывания кластеров исследованы методы Уорда, ближайшего и дальнего соседей, средней связи, Маккуити и центроидный.

Для оценки точности отнесения транксриптов РНК к заданному гену используется мера

$$A = 100 \cdot (N_1 + N_2)/N, \tag{4}$$

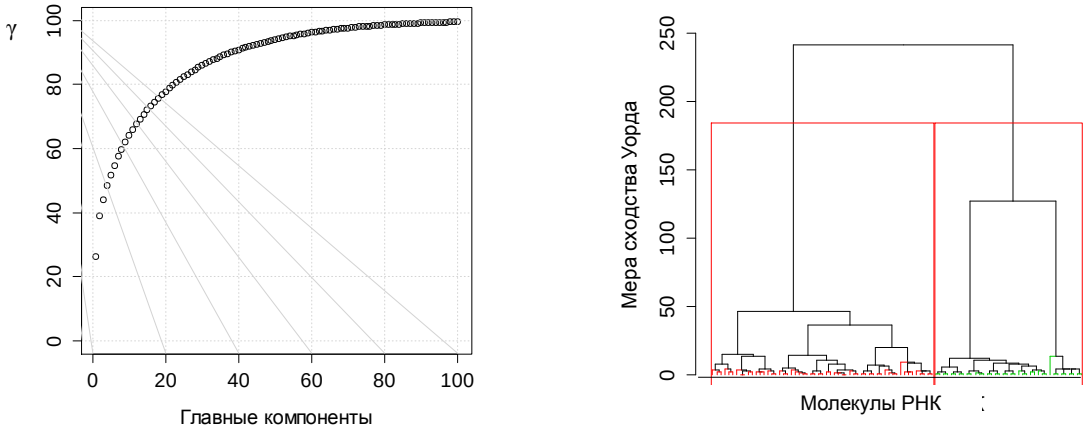
где N_1 и N_2 – числа правильно классифицированных транскриптов для двух генов; N – общее число транскриптов двух генов.

Результаты исследования. Сравнительный анализ РНК-продуктов нормальных генов человека. В ходе анализа полного набора признаков экзонов 10 пар модельных генов с использованием метода главных компонент было отобрано от 50 до 100 наиболее значимых компонент, объясняющих 99 % вариаций в данных. Пример, демонстрирующий результаты работы метода главных компонент, представлен на рис. 3, *а*. Применение метода снижения размерности данных позволило в 15–30 раз сократить пространство признаков экзонов, что привело почти к пропорциональному увеличению производительности вычислений в ходе последующего анализа.



а) *б)* *в)* *г)*

Рис. 2. Примеры оценки качества неудовлетворительного (*а*, *б*) и успешного (*в*, *г*) иерархического кластерного анализа экзонов модельных генов: *а*) и *в*) дендрограммы иерархического кластерного анализа (красным цветом выделены кластеры сходных экзонов); *б*) и *г*) гистограммы распределений числа экзонов в кластерах (вертикальные полосы (np_j , j = 1, 2, …, 26) – нормированные вероятности для равномерной дискретной случайной величины, распределенной в интервале [1; 26])



а) *б)*

Рис. 3. Результаты работы методов главных компонент и иерархического кластерного анализа для пары генов MUC20-OT1 и SOX2-OT: *а*) относительная кумулятивная доля разброса γ для первых 100 главных компонент, полученная в результате применения метода главных компонент для анализа экзонов пары генов; *б*) результаты работы алгоритма иерархической кластеризации пула уникальных экспериментальных транскриптов двух модельных генов

Результаты иерархического кластерного анализа экзонов 10 пар модельных генов с использованием различных способов вычисления расстояний и связывания кластеров представлены в табл. 2. Наилучшие результаты этого анализа характерны для связывания кластеров по Уорду, для которых среднее качество кластерного анализа <Q> выше 1,3. Наихудшими являются ме-

тоды, включающие медианное и ближайшего соседа объединения кластеров. Следует отметить, что выбор расстояния не столь критичен. Тем не менее иерархический кластерный анализ с использованием расстояний Минковского и связывания по Уорду является наиболее эффективным способом группировки экзонов генов и для дальнейшего исследования выбраны результаты именно этого анализа.

Таблица 2

Рейтинг способов вычисления расстояний и методов связывания иерархического кластерного анализа экзонов 10 пар модельных генов (<к>, <χ^2> и <*Q*> – оценки математического ожидания для величин к, χ^2 и *Q*)

Ранг	Расстояние	Метод	<к>	<χ^2>	<*Q*>
1	minkowski	ward	0,543	59,733	1,311
2	maximum	ward	0,556	61,422	1,294
3	manhattan	ward	0,304	64,557	1,180
4	euclidean	ward	0,478	72,112	1,117
5	maximum	complete	0,573	81,229	0,998
6	minkowski	complete	0,605	135,909	0,711
7	euclidean	complete	0,625	258,160	0,567
8	maximum	average	0,687	304,849	0,542
9	maximum	mcquitty	0,619	234,470	0,542
10	minkowski	average	0,749	436,950	0,517
11	minkowski	mcquitty	0,662	381,170	0,496
12	euclidean	mcquitty	0,707	579,726	0,471
13	euclidean	average	0,820	1154,007	0,467
14	manhattan	average	0,854	2411,358	0,450
15	manhattan	centroid	0,831	2678,246	0,436
16	manhattan	complete	0,598	598,123	0,420
17	manhattan	single	0,797	2599,807	0,419
18	manhattan	median	0,774	2662,043	0,407
19	manhattan	mcquitty	0,715	1219,003	0,407
20	euclidean	centroid	0,710	2566,069	0,376
21	euclidean	single	0,701	2248,179	0,376
22	maximum	centroid	0,644	1425,691	0,371
23	minkowski	centroid	0,655	1757,906	0,365
24	euclidean	median	0,629	2560,567	0,336
25	minkowski	single	0,566	1900,626	0,314
26	minkowski	median	0,536	2092,882	0,295
27	maximum	single	0,512	1776,142	0,292
28	maximum	median	0,469	1555,219	0,274

Результаты иерархического кластерного анализа транскриптов 10 пар различных генов с использованием различных способов вычисления расстояний и связывания кластеров представлены в табл. 3. Пример успешного разделения экспериментально подтвержденных транскриптов показан на рис. 3, *б*. По вертикальной оси указано расстояние Уорда, по горизонтальной – молекулы РНК. Контурами выделено итоговое разбиение транскриптов РНК на два кластера, соответствующих генам MUC20-OT1 и SOX2-OT. Наилучшим способом

объединения кластеров является связывание по Уорду, наилучшими расстояниями – Жаккара, косинусное и Джаро – Винклера. Расстояние Жаккара и связывание кластеров по Уорду обеспечивают наиболее высокую точность кластеризации. Средняя точность разделения составляет 99,5 % для рассмотренных пар модельных генов, что подтверждает работоспособность разработанного подхода: транскрипты пар модельных генов надежно разделяются на два класса. Следует отметить, что анализ пар генов, характеризуемых полным набором экзонов, не позволяет достичь точности разделения более 88–89 %. Это является весомым аргументом в пользу обязательного применения метода главных компонент не только для снижения размерности данных, но и для повышения точности вычислений.

Таблица 3

Рейтинг способов вычисления расстояний и методов связывания иерархического кластерного анализа транскриптов 10 пар модельных генов (<*A*> и с.к.о. – оценки математического ожидания и среднеквадратического отклонения для величины точности классификации *A*)

Ранг	Расстояние	Метод	<*A*>	С.к.о.
1	jaccard	ward	99,5	0,9
2	cosine	ward	99,1	1,4
3	jw	ward	95,5	6,6
4	lcs	ward	89,2	17,2
5	qgram	ward	88,0	18,1
6	jaccard	single	83,1	22,0
7	cosine	single	78,3	23,0
8	lcs	complete	71,5	22,5
9	qgram	complete	67,5	23,7
10	cosine	complete	67,2	10,8
11	jw	single	66,9	24,4
12	osa	ward	65,6	17,7
13	lv	ward	65,6	17,7
14	dl	ward	65,6	17,7
15	jaccard	complete	65,6	13,9
16	lv	complete	61,2	14,6
17	osa	complete	61,0	14,3
18	dl	complete	61,0	14,3
19	jw	complete	60,3	8,8
20	qgram	single	59,6	14,3
21	lcs	single	56,6	12,3
22	dl	single	56,4	12,0
23	soundex	ward.D2	55,4	10,9
24	soundex	single	54,7	8,9
25	soundex	complete	54,7	8,9
26	osa	single	54,5	11,5
27	lv	single	54,5	11,5

Сравнительный анализ экспериментально подтвержденных и предсказываемых РНК-продуктов гибридного онкогена RUNX1/RUNX1T1 человека. Разделение экспериментальных и теоретически возможных транскриптов гибридного онкогена RUNX1/RUNX1T1 яв-

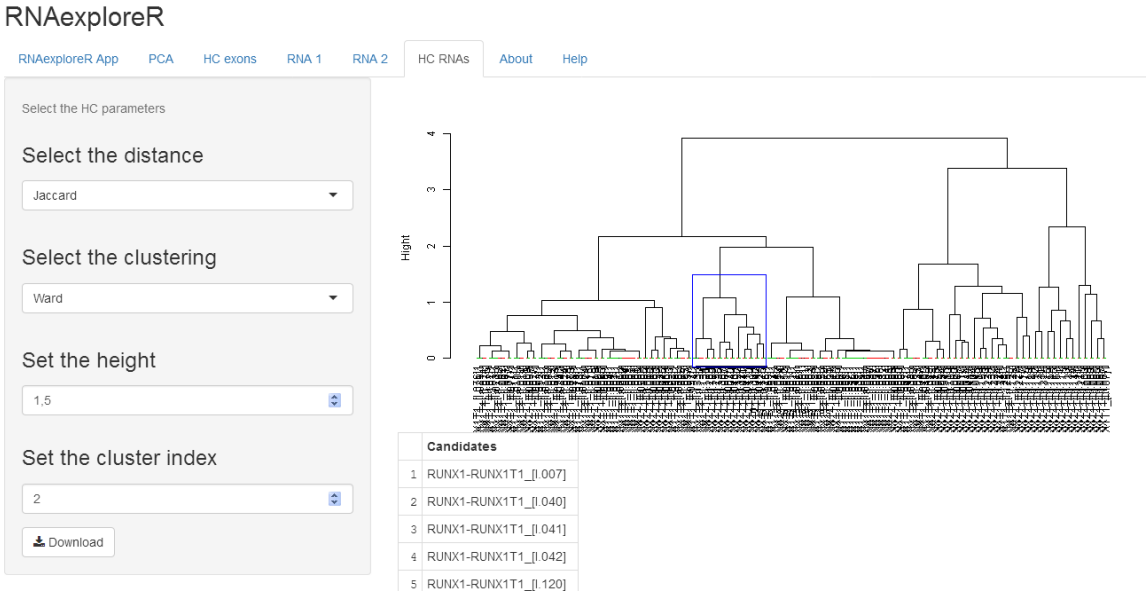
ляется более сложной задачей, чем классификация транскриптов модельных генов. Это обусловлено несколькими причинами. Во-первых, предсказываемые изоформы РНК, будучи путями графа сплайсинга, реконструированного по реальным транскриптам, состоят из тех же экзонов, что и экспериментально обнаруженные молекулы РНК. Во-вторых, многие альтернативные экзоны гибридного онкогена RUNX1/RUNX1T1 лишь незначительно отличаются друг от друга по нуклеотидной последовательности (образуются на основе канонических экзонов путем применения альтернативных 5’ и (или) 3’ сайтов сплайсинга), что делает их плохо различимыми или полностью неразличимыми по тем признакам, которые были использованы для описания экзонов. В-третьих, многие пути в графе сплайсинга уникальны лишь по таким упомянутым выше и плохо различимым экзонам. Как следствие, и сами пути очень плохо различаются. Наконец, в-четвертых, как было показано ранее [5, 6, 33], одной из ключевых мод альтернативного сплайсинга РНК изучаемого онкогена является выбор альтернативных 5’ и (или) 3’ сайтов сплайсинга, находящихся, как правило, на небольшом расстоянии от канонических сайтов. Наравне с пропуском экзонов эта мода сплайсинга обеспечивает образование подавляющего большинства изоформ РНК гибридного онкогена RUNX1/RUNX1T1, что опять же затрудняет разделение таких вариантов РНК.

Тем не менее, несмотря на отмеченные выше сложности, разработанный алгоритм анализа обладает достаточной чувствительностью к небольшим различиям в структуре транскриптов, чтобы относительно успешно их разделять. Поскольку транскрипты гибридного онкогена RUNX1/RUNX1T1 генерируются разными путями и между ними есть существенные структурные различия [6, 33], то первоначально была проведена кластеризация самих экспериментально подтвержденных транскриптов и только после этого сопоставлены эти транскрипты с теоретически предсказанными изоформами РНК. Такой подход вместе с разработанным алгоритмом анализа позволяет разнести экспериментально подтвержденные и предсказанные транскрипты по двум разным группам, что наглядно подтверждает работоспособность разработанного метода (рис. 4).

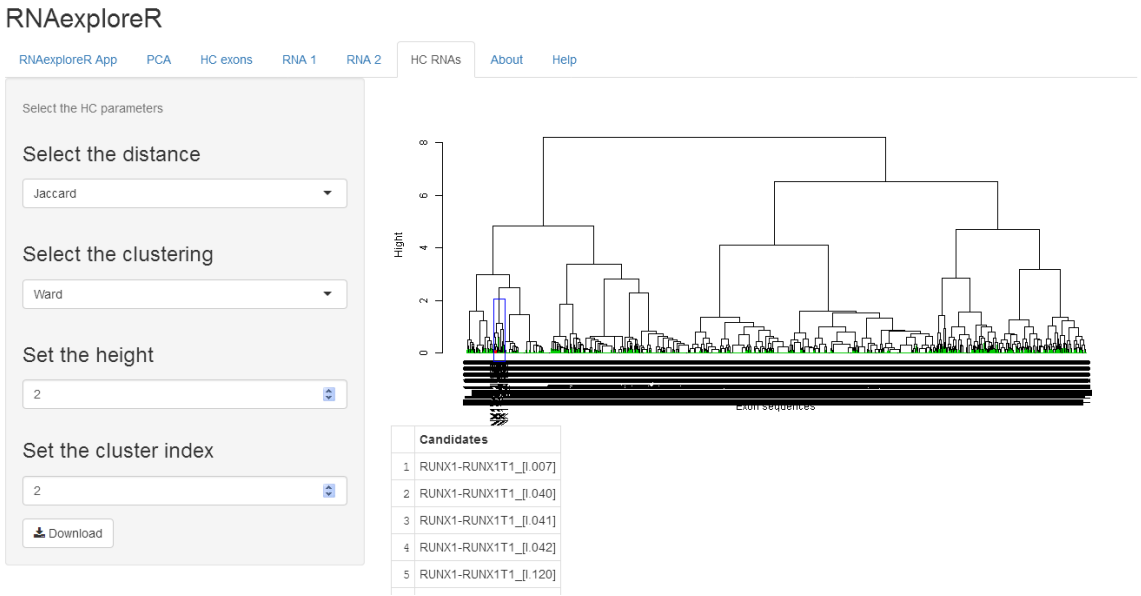
Следует отметить, что в данной работе определение изоформ РНК эквивалентно нахождению наиболее вероятных транкриптов из списка теоретически предсказанных молекул РНК, сформированного путем полного обхода графа сплайсинга. В случае добавления некоторой новой изоформы, не входящей в список предсказанных изоформ данного гена, исследование выполняется аналогично: формируются метки экзонов данной изоформы, вычисляются расстояния от изоформы до экспериментально подтвержденных транскриптов РНК, определяется минимальное расстояние до некоторого экспериментально подтвержденного транскрипта РНК, мера близости к которому является предсказательной оценкой событий альтернативного сплайсинга.

Программный пакет RNAexploreR. Для реализации программного обеспечения используются различные вычислительные платформы и технологии программирования. В большинстве публикаций, посвященных сравнительному анализу бесплатных пакетов по дисциплине «Интеллектуальный анализ данных», нет явного лидера. В настоящее время в открытом доступе находится большое количество программных средств интеллектуального анализа данных, среди которых можно выделить WEKA, Tanagra, Rapid Miner, KNIME, Python- и R-платформы [34]. Достоинствами того или иного программного ресурса являются: вычислительная производительность, широкий набор подключаемых библиотек, кроссплатформенность, возможность выполнения параллельных вычислений и работы напрямую с базами и хранилищами данных.

Основным преимуществом среды статистического программирования R является возможность использования огромного набора биоинформационных алгоритмов, алгоритмов интеллектуального анализа данных и разнообразных статистических вычислительных ресурсов научного сообщества [35, 36]. Главный ее недостаток – невысокая вычислительная производительность. Данное ограничение можно частично или полностью устранить с помощью процедур распараллеливания вычислений, подключения библиотек высокопроизводительных математических вычислений (например, Microsoft R Open (MRO) и Intel Math Kernel Library (MKL)) или же процедур компиляции и подключения программного кода, реализованного на других специализированных языках программирования (например, с помощью R-пакетов Rcpp, Rinside, inline, rJava, reticulate). Наиболее популярными пакетами для разработки пользовательских интерфейсов программных приложений, интегрирующими R-коды, являются gWidgets, rpanel, svDialogs, RGtk2, qtbase, tcltk [35].



а)



б)

Рис. 4. Дендрограммы иерархического кластерного анализа: *а)* уникальных экспериментально подтвержденных транскриптов гибридного онкогена RUNX1/RUNX1T1; *б)* случайной выборки из 3000 предсказываемых изоформ РНК и кластера экспериментально подтвержденных транскриптов (выделен рамкой синего цвета). По вертикальной оси указано расстояние Уорда, по горизонтальной – молекулы РНК. Мера расчета близости молекул – расстояние Жаккара

Новое направление в разработке R-приложений связано с созданием «реактивных» веб-интерфейсов с помощью пакета Shiny и размещением программной реализации на ресурсе shinyapps.io, предоставляемом разработчиками открытого программного обеспечения RStudio [37]. Достоинством данного подхода является возможность удаленной работы с приложением широкой научной аудитории пользователей в режиме онлайн через сеть Интернет. Для реализации программного приложения в работе выбраны вычислительная среда R и пакет Shiny для создания веб-интерфейса приложения.

Рассмотренные алгоритмы запрограммированы на языке R и собраны в единый модуль. Реализовано несколько вариантов программных пакетов: онлайн, офлайн, оптимизированный и распараллеленный. Онлайн-вариант программного пакета размещен в сети Интернет на веб-ресурсе https://dsa-cm.shinyapps.io/NIR_bio_code_Sh-MolBio/ [38].

Разработанное веб-приложение RNAexploreR интегрирует реализованные алгоритмы методики вычислительного подхода. Главное окно интерфейса пакета (рис. 5) состоит из восьми панелей, соответствующих пяти этапам анализа, информации об авторах разработки и странице помощи. На каждом этапе анализа пользователь должен выбрать соответствующий файл (экзонов, экспериментально подтвержденных транскриптов или молекулы РНК 1 / RNA 1, теоретически предсказанных транскриптов или молекулы РНК 2 / RNA 2) и установить системные параметры алгоритмов. Результаты анализа, а также список транскриптов сохраняются в отдельный csv-файл.

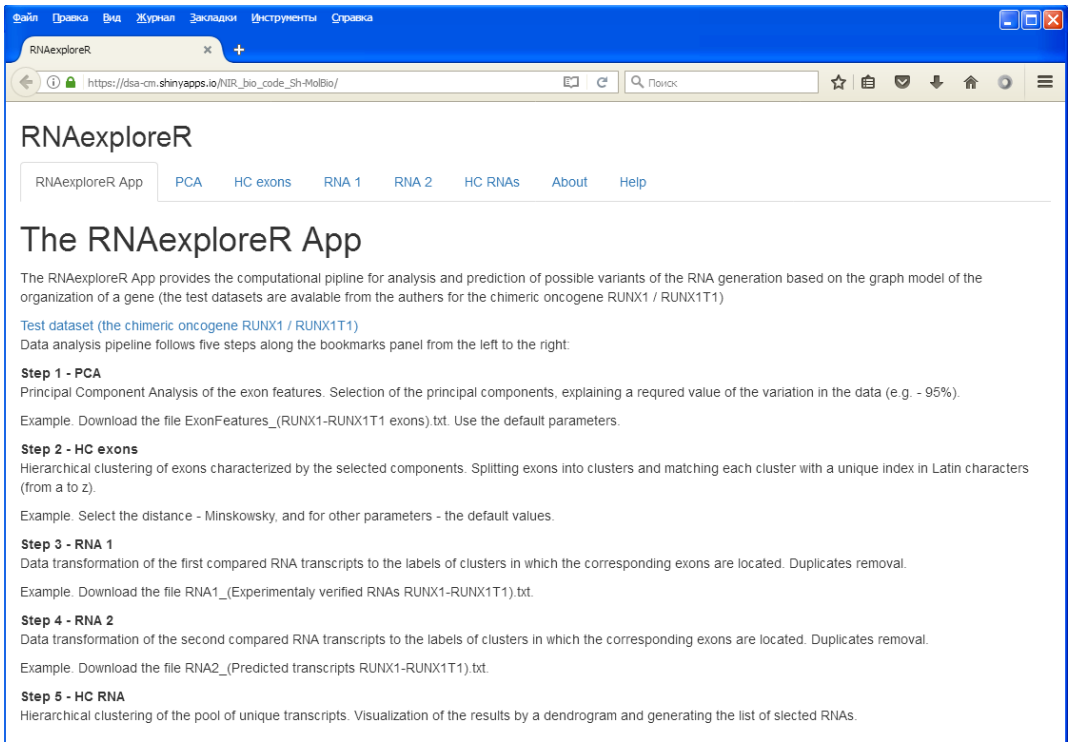


Рис. 5. Главное окно интерфейса пакета RNAexploreR

Заключение. Впервые предложен и исследован на модельных генах человека системный вычислительный подход к сравнению транскриптов генов человека, основанный на применении алгоритмов снижения размерности данных, иерархического кластерного анализа, сравнения символьных последовательностей. Средняя точность отнесения транскрипта к заданному гену составляет более 99 % для рассмотренных пар нормальных генов человека. Вычислительный подход позволяет разнести выборки экспериментально подтвержденных и предсказанных транскриптов по двум разным группам молекул РНК гибридного онкогена RUNX1/RUNX1T1.

Разработаны программный пакет RNAexploreR и одноименное веб-приложение, интегрирующие все реализованные алгоритмы анализа альтернативного сплайсинга РНК-продуктов генов человека. Новизна разработки состоит в том, что пакет включает методологию анализа событий альтернативного сплайсинга молекул РНК генов человека, основанную на моделировании экзонных графов генов и применении эффективных алгоритмов интеллектуального анализа данных. Разработка имеет ряд преимуществ в сравнении с традиционными программными решениями: позволяет осуществлять высокоточную классификацию молекул РНК генов человека по множеству экзонных признаков и структуре таких молекул, автоматизирует анализ генетических данных, обеспечивает высокое качество анализа в условиях высокого экспери-

ментального шума, интегрирует набор усовершенствованных алгоритмов интеллектуального анализа данных, предоставляет возможность удаленной работы с приложением широкой научной аудитории пользователей в режиме онлайн через сеть Интернет, использует распараллеливание вычислительных ресурсов.

Предложенные алгоритмы и программное обеспечение могут применяться для изучения организации и функционирования как аберрантных, так и нормальных генов человека, а получаемые при этом данные могут быть полезны для дифференциальной диагностики и построения прогноза течения заболеваний, имеющих генетическую природу.

Список использованных источников

1. Baralle, F. E. Alternative splicing as a regulator of development and tissue identity / F. E. Baralle, J. Giudice // Nat. Rev. Mol. Cell Biol. – 2017. – Vol. 18. – P. 437–451.
2. Nilsen, T. W. Expansion of the eukaryotic proteome by alternative splicing / T. W. Nilsen, B. R. Graveley // Nature. – 2010. – Vol. 463. – P. 457–463.
3. Ramanouskaya, T. V. The determinants of alternative RNA splicing in human cells / T. V. Ramanouskaya, V. V. Grinev // Mol. Genet. Genomics. – 2015. – Vol. 292. – P. 1175–1195.
4. Sequence, structure, and context preferences of human RNA binding proteins / D. Dominguez [et al.] // Mol. Cell. – 2018. – Vol. 70. – P. 854–867.
5. Изучение закономерностей сплайсинга РНК гибридного онкогена RUNX1-RUNX1T1 человека с помощью методов интеллектуального анализа данных и высокопроизводительного секвенирования / И. Н. Ильюшёнок [и др.] // Молекулярная и прикладная генетика. – 2017. – № 23. – С. 92–101.
6. Decoding of exon splicing patterns in the human RUNX1-RUNX1T1 fusion gene / V. V. Grinev [et al.] // Intern. J. Biochem. Cell Biol. – 2015. – Vol. 68. – P. 48–58.
7. Deciphering the splicing code / Y. Barash [et al.] // Nature. – 2010. – Vol. 465. – P. 53–59.
8. Расширяя гипотезу «двух ударов»: молекулярные механизмы RUNX1-RUNX1T1-опосредованного лейкозогенеза / И. Н. Ильюшёнок [и др.] // Журн. Белорус. гос. ун-та. Биология. – 2017. – № 2. – С. 3–16.
9. Ensembl 2018 / D. R. Zerbino [et al.] // Nucleic Acids Res. – 2018. – Vol. 46(D1). – P. D754–D761.
10. Яцков, Н. Н. Интеллектуальный анализ данных : пособие / Н. Н. Яцков. – Минск : БГУ, 2014. – 151 с.
11. Bramer, M. Principles of Data Mining / M. Bramer. – 2nd ed. – London : Springer, 2013. – 440 p.
12. Aggarwal, C. C. Data Mining: The Textbook / C. C. Aggarwal. – Gewerbestrasse : Springer, 2015. – 734 p.
13. Hastie, T. The Elements of Statistical Learning. Data Mining, Inference, and Prediction / T. Hastie, R. Tibshirani, J. Friedman. – 2nd ed. – N. Y. : Springer, 2009. – 739 p.
14. Разработка алгоритмов и программных средств классификации кодирующих и некодирующих нуклеотидных последовательностей / В. Р. Закирова [и др.] // Информатика. – 2019. – Т. 16, № 2. – С. 111–120.
15. Zhang, S. W. Gene prediction in metagenomic fragments with deep learning / S. W. Zhang, X. Y. Jin, T. Zhang // BioMed Res. Intern. – November 2017.
16. Al-Ajlan, A. Feature selection for gene prediction in metagenomic fragments / A. Al-Ajlan, A. El Allali // BioData Min. – 2018. – Vol. 11.
17. Al-Ajlan, A. CNN-MGP: Convolutional neural networks for metagenomics gene prediction / A. Al-Ajlan, A. El Allali // Interdisciplinary Sciences. – December 2018.
18. Прикладная статистика: классификация и снижение размерности : справ. изд. / С. А. Айвазян [и др.] ; под ред. С. А. Айвазяна. – М. : Финансы и статистика, 1989. – 607 с.
19. Jolliffie, I. T. Principal Component Analysis / I. T. Jolliffie. – 2nd ed. – N. Y. : Springer, 2002. – 487 p.
20. Hyvaerinen, A. Independent Component Analysis / A. Hyvaerinen, J. Karhunen, O. Erkki. – N. Y. : John Wiley&Sons Inc., 2001. – 481 p.
21. Лагутин, М. Б. Наглядная математическая статистика : учеб. пособие / М. Б. Лагутин. – М. : БИНОМ. Лаборатория знаний, 2007. – 472 с.
22. Saeys, Y. A review of feature selection techniques in bioinformatics / Y. Saeys, I. Inza, P. Larranaga // Bioinformatics. – 2007. – Vol. 23. – P. 2507–2517.
23. Волков, А. В. Отбор информативных признаков экзонов генов человека / А. В. Волков, Н. Н. Яцков, В. В. Гринев // Журн. Белорус. гос. ун-та. Математика. Информатика. – 2019. – № 1. – С. 77–89.
24. Мандель, И. Д. Кластерный анализ / И. Д. Мандель. – М. : Финансы и статистика, 1988. – 176 с.
25. Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP / А. А. Барсегян [и др.]. – 2-е изд. – СПб. : БХВ-Петербург, 2007. – 384 с.
26. Леск, А. М. Введение в биоинформатику : пер. с англ. / А. М. Леск. – М. : БИНОМ. Лаборатория знаний, 2009. – 318 с.

27. Van der Loo, M. P. J. The stringdist package for approximate string matching / M. P. J. Van der Loo // The R Journal. – 2014. – Vol. 6. – P. 111–122.

28. Uragun, B. The discrimination of interaural level difference sensitivity functions: development of a taxonomic data template for modeling / B. Uragun, R. Rajan // BMC Neuroscience. – 2013. – Vol. 14. – P. 1–19.

29. Yatskou, M. Computer Simulation of Energy Relaxation and Transport in Organized Porphyrin Systems / M. Yatskou . – The Netherlands, Wageningen : Ponsen & Looijen Printing Establishment, 2001. – 176 p.

30. Boytsov, L. Indexing methods for approximate dictionary searching: comparative analyses / L. Boytsov // ACM Journal of Experimental Algorithmics. – 2011. – Vol. 16. – P. 1–88.

31. Navarro, G. A guided tour to approximate string matching / G. Navarro // ACM Computing Surveys. – 2001. – Vol. 33. – P. 31–88.

32. Cohen, W. A comparison of string metrics for matching names and records / W. Cohen // KDD. – 2003. – Vol. 3. – P. 73–78.

33. Вклад различных механизмов генерации альтернативных транскриптов в разнообразие мРНК гибридного онкогена RUNX1-RUNX1T1 человека / И. Н. Ильюшёнок [и др.] // Журн. Белорус. гос. ун-та. Биология. – 2019. – № 2. – С. 45–59.

34. Программный пакет RNAexploreR для предсказания вариантов альтернативного сплайсинга в первичных мРНК химерного онкогена RUNX1/RUNX1T1 человека / Н. Н. Яцков [и др.] // Информационные технологии и системы 2018 (ИТС–2018) : материалы Междунар. науч. конф., Минск, 25 окт. 2018 г. – Минск : БГУИР, 2018. – С. 282–283.

35. R Core Team. R: A language and Environment for Statistical Computing [Electronic recourse] / R Foundation for Statistical Computing. – 2014. – Mode of access: http://www.R-project.org/. – Date of access: 08.02.2019.

36. Gentleman, R. Bioconductor: Open software development for computational biology and bioinformatics / R. Gentleman, V. J. Carey, D. M. Bates // Genome Biology. – 2004. – Vol. 5, no. 10, R80.

37. RStudio: Integrated Development for R [Electronic recourse]. – 2015. – Mode of access: http://www.rstudio.com/. – Date of access: 13.06.2019.

38. Yatskou, M. M. RNAexplorerR : Application of the Computational Pipline for Analysis and Prediction of Possible Variants of the RNA Generation Based on the Graph Model of the Organization of a Gene [Electronic recourse] / M. M. Yatskou, V. V. Skakun, V. V. Grinev. – Mode of access: https://dsa-cm.shinyapps.io/NIR_bio_code_Sh-MolBio/. – Date of access: 13.06.2019.

References

1. Baralle F. E., Giudice J. Alternative splicing as a regulator of development and tissue identity. *Nature Reviews Molecular Cell Biology*, 2017, vol. 18, pp. 437–451.

2. Nilsen T. W., Graveley B. R. Expansion of the eukaryotic proteome by alternative splicing. *Nature*, 2010, vol. 463, pp. 457–463.

3. Ramanouskaya T. V., Grinev V. V. The determinants of alternative RNA splicing in human cells. *Molecular Genetics and Genomics*, 2015, vol. 292, pp. 1175–1195.

4. Dominguez D., Freese P., Alexis M. S., Su A., Hochman M., …, Burge C. B. Sequence, structure, and context preferences of human RNA binding proteins. *Molecular Cell*, 2018, vol. 70, pp. 854–867.

5. Ilyushonak I. M., Gunko E. P., Antonovich M. L., Yatskou M. M., Kustanovich A. M., ..., Grinev V. V. Izuchenie zakonomernostej splajsinga RNK gibridnogo onkogena RUNX1-RUNX1T1 cheloveka s pomoshyu metodov intellektualnogo analiza dannyh i vysokoproizvoditelnogo sekvenirovaniya [Study of RNA splicing patterns of the human RUNX1-RUNX1T1 fusion oncogene by the methods of data mining and high-throughput DNA sequencing]. Molekuljarnaja i prikladnaja genetika [*Molecular and Applied Genetics*], 2017, vol. 23, pp. 92–101 (in Russian).

6. Grinev V. V., Migas A. A., Kirsanava A. D., Mishkova O. A., Siomava N., ..., Aleinikova O. V. Decoding of exon splicing patterns in the human RUNX1-RUNX1T1 fusion gene. *The International Journal of Biochemistry & Cell Biology*, 2015, vol. 68, pp. 48–58.

7. Barash Y., Calarco J. A., Gao W., Pan Q., Wang X., ..., Frey B. J. Deciphering the splicing code. *Nature*, 2010, vol. 465, pp. 53–59.

8. Ilyushonak I. M., Saurytskaya H. A., Yatskou M. M., Skakun V. V., Grinev V. V. Rasshiryaya gipotezu "dvuh udarov": molekulyarnye mehanizmy RUNX1-RUNX1T1-oposredovannogo lejkozogeneza [Extending the "two-hits" hypothesis: the molecular mechanisms of RUNX1-RUNX1T1-mediated leukemogenesis]. Zhurnal Belorusskogo gosudarstvennogo universiteta. Biologija [*Journal of the Belarusian State University. Biology*], 2017, no. 2, pp. 3–16 (in Russian).

9. Zerbino D. R., Achuthan P., Akanni W., Amode M. R., Barrell D., ..., Flicek P. Ensembl 2018. *Nucleic Acids Research*, 2018, vol. 46(D1), pp. D754–D761.

10. Yatskou M. M. Intellektualnyj analiz dannyh. *Data Mining*. Minsk, Belarusian State University, 2014, 151 p. (in Russian).

11. Bramer M. *Principles of Data Mining*. 2nd ed. London, Springer, 2013, 440 p.

12. Aggarwal C. C. *Data Mining: The Textbook*. Gewerbestrasse, Springer, 2015, 734 p.

13. Hastie T., Tibshirani R., Friedman J. *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*. 2nd ed. New York, Springer, 2009, 739 p.

14. Zakirava V. R., Syrokvash D. A., Hileuski S. V., Nazarov P. V., Yatskou M. M. Razrabotka algoritmov i programmnyh sredstv klassifikacii kodiruyushih i nekodiruyushih nukleotidnyh posledovatelnostej [Development of algorithms and software for classification of nucleotide sequences]. Informatika [*Informatics*], 2019, vol. 16, no. 2, pp. 111–120 (in Russian).

15. Zhang S. W., Jin X. Y., Zhang T. Gene prediction in metagenomic fragments with deep learning. *BioMed Research International*, November 2017. DOI: 10.1155/2017/4740354

16. Al-Ajlan A., El Allali A. Feature selection for gene prediction in metagenomic fragments. *BioData Mining*, 2018, vol. 11. DOI: 10.1186/s13040-018-0170-z

17. Al-Ajlan A., El Allali A. CNN-MGP: Convolutional neural networks for metagenomics gene prediction. *Interdisciplinary Sciences*, December 2018. DOI: 10.1007/s12539-018-0313-4

18. Aivazyan S. A., Buchstaber V. M., Yenyukov I. S., Meshalkin L. Prikladnaya statistika: klassifikaciya i snizhenie razmernosti. *Applied Statistics: Classification and Reduction of Dimensionality.* In S. A. Aivazyan (ed.). Moscow, Finansy i statistika, 1989, 607 p. (in Russian).

19. Jolliffie I. T. *Principal Component Analysis*. 2nd ed. New York, Springer, 2002, 487 p.

20. Hyvaerinen A., Karhunen J., Erkki O. *Independent Component Analysis*. New York, John Wiley&Sons Inc., 2001, 481 p.

21. Lagutin, M. B. Naglyadnaya matematicheskaya statistika. *Visual Mathematical Statistics*. Moscow, BINOM, Laboratoriya znanij, 2007, 472 p. (in Russian).

22. Saeys Y., Inza I., Larranaga P. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 2007, vol. 23, pp. 2507–2517.

23. Volkau A. U., Yatskou M. M., Grinev V. V. Otbor informativnyh priznakov ekzonov genov cheloveka [Selecting informative features of human gene exons]. Zhurnal Belorusskogo gosudarstvennogo universiteta. Matematika. Informatika [*Journal of the Belarusian State University. Mathematics and Informatics*], 2019, no. 1, pp. 77–89 (in Russian).

24. Mandel I. D. Klasternyj analiz. *Cluster Analysis*. Moscow, Finansy i statistika, 1988, 176 p. (in Russian).

25. Barsegyan A. A., Kupriyanov M. S., Stepanenko V. V., Holod I. I. Tehnologii analiza dannyh : Data Mining, Visual Mining, Text Mining, OLAP. *Data Analysis Technologies : Data Mining, Visual Mining, Text Mining, OLAP*. 2nd ed. Saint Petersburg, BHV-Peterburg, 2007, 384 p.

26. Lesk A. M. *Introduction to Bioinformatics*. Oxford, Oxford University Press, 2002, 283 p.

27. Van der Loo M. P. J. The stringdist package for approximate string matching. *The R Journal*, 2014, vol. 6, pp. 111–122.

28. Uragun B., Rajan R. The discrimination of interaural level difference sensitivity functions: development of a taxonomic data template for modeling. *BMC Neuroscience*, 2013, vol.14, pp. 1–19.

29. Yatskou M. *Computer Simulation of Energy Relaxation and Transport in Organized Porphyrin Systems*. The Netherlands, Wageningen, Ponsen & Looijen Printing Establishment, 2001, 176 p.

30. Boytsov L. Indexing methods for approximate dictionary searching: comparative analyses. *ACM Journal of Experimental Algorithmics*, 2011, vol. 16, pp. 1–88.

31. Navarro G. A guided tour to approximate string matching. *ACM Computing Surveys*, 2001, vol. 33, pp. 31–88.

32. Cohen W. A comparison of string metrics for matching names and records. *KDD*, 2003, vol. 3, pp. 73–78.

33. Ilyushonak I. M., Migas A. A., Sukhareuski A. Y., Schneider A. D., Grinev V. V. Vklad razlichnyh mehanizmov generacii alternativnyh transkriptov v raznoobrazie mRNK gibridnogo onkogena RUNX1-RUNX1T1 cheloveka [The contribution of various mechanisms to mRNA diversity of human fusion oncogene RUNX1-RUNX1T1]. Zhurnal Belorusskogo gosudarstvennogo universiteta. Biologija [*Journal of the Belarusian State University. Biology*], 2019, no. 2, pp. 45–59 (in Russian).

34. Yatskou M. M., Skakun V. V., Grinev V. V. Programmnyj paket RNAexploreR dlya predskazaniya variantov alternativnogo splajsinga v pervichnyh mRNK himernogo onkogena RUNX1/RUNX1T1 cheloveka [The software package RNAexploreR for predicting alternative splicing variants in primary mRNAs of the human chimeric oncogen RUNX1/RUNX1T1]. Informacionnye tehnologii i sistemy 2018 (ITS–2018): materialy Mezhdunarodnoj nauchnoj konferencii, Minsk, 25 oktjabrja 2018 [*Information Technologies and*

Systems 2018 (ITS–2018): Proceedings of the International Scientific Conference, Minsk, 25 October 2018]. Minsk, Belorusskij gosudarstvennyj universitet informatiki i radioelektroniki, 2018, pp. 282–283 (in Russian).

35. *R Core Team. R: A language and Environment for Statistical Computing*, 2014. Available at: http://www.R-project.org/ (accessed 08.02.2019).

36. Gentleman R., Carey V. J., Bates D. M. Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, 2004, vol. 5, no. 10, R80.

37. *RStudio: Integrated Development for R*, 2015. Available at: http://www.rstudio.com/ (accessed 13.06.2019).

38. Yatskou M. M., Skakun V. V., Grinev V. V. *RNAexplorerR : Application of the Computational Pipline for Analysis and Prediction of Possible Variants of the RNA Generation Based on the Graph Model of the Organization of a Gene*. Available at: https://dsa-cm.shinyapps.io/NIR_bio_code_Sh-MolBio/ (accessed 13.06.2019).

Информация об авторах

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

Скакун Виктор Васильевич, кандидат физико-математических наук, доцент, заведующий кафедрой системного анализа и компьютерного моделирования, факультет радиофизики и компьютерных технологий, Белорусский государственный университет, Минск, Беларусь.
E-mail: skakun@bsu.by

Гринев Василий Викторович, кандидат биологических наук, доцент, доцент кафедры генетики, биологический факультет, Белорусский государственный университет, Минск, Беларусь.
E-mail: grinev_vv@bsu.by

Information about the authors

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

Victor V. Skakun, Cand. Sci. (Phys.-Math.), Associate Professor, Head of Department of Systems Analysis and Computer Modelling, Faculty of Radiophysics and Computer Technologies, Belarusian State University, Minsk, Belarus.
E-mail: skakun@bsu.by

Vasily V. Grinev, Cand. Sci. (Biol.), Associate Professor, Department of Genetics, Faculty of Biology, Belarusian State University, Minsk, Belarus.
E-mail: grinev_vv@bsu.by