СТАТИСТИЧЕСКИЙ ПОДХОД К ПРЕДСКАЗАНИЮ СОБЫТИЙ АЛЬТЕРНАТИВНОГО СПЛАЙСИНГА В ПЕРВИЧНЫХ МРНК ГЕНОВ ЧЕЛОВЕКА

Яцков Н. Н., Скакун В. В., Гринев В. В.
Кафедра системного анализа и компьютерного моделирования, кафедра генетики, Белорусский государственный университет
Минск, Республика Беларусь
E-mail: {yatskou, skakun}@bsu.by

Разработан статистический подход к предсказанию событий альтернативного сплайсинга в первичных мРНК генов человека. На основе исследованных алгоритмов снижения размерности данных, иерархического кластерного анализа, вычисления расстояний между символьными последовательностями выполнено сравнение экспериментально подтвержденных транскриптов модельных генов человека. Точность предсказания разработанного подхода составляет 90-95% для рассмотренных пар модельных генов.

Введение

Конститутивный и альтернативный сплайсинг является фундаментальным процессом, протекающим во всех без исключения клетках эукариот и приводящим к образованию зрелых функциональных РНК-продуктов [1]. Однако, принципы (правила) комбинаторики экзонов во время сплайсинга до сих пор не установлены [2]. Следует отметить об ограниченном применении или даже полном отсутствии стандартов или единых систематизированных статистических подходов к анализу и интерпретации возможных экзонных последовательностей генов человека. Для выяснения принципов, по которым идет комбинаторика экзонов во время сплайсинга [3], требуется разработка статистических алгоритмов и программных средств для анализа и предсказания разнообразных вариантов генерации РНК.

Целью работы является разработка статистического системного подхода к анализу и предсказанию событий альтернативного сплайсинга в первичных мРНК генов человека. Реализованы и исследованы наиболее эффективные алгоритмы интеллектуального анализа данных. Проверка работоспособности разработанных алгоритмов выполнена на примере набора модельных генов человека [4].

I. Объект и предмет исследования

Объект исследования – альтернативный сплайсинг РНК онкогена человека. В качестве примера рассмотрен набор из 14 произвольно отобранных негомологичных генов человека [4]. Для каждого из рассмотренных генов из базы данных Ensembl получены списки уникальных экзонов и экспериментально подтвержденных транскриптов. Предметом исследования являются алгоритмы интеллектуального анализа данных, позволяющие предсказать события альтернативного сплайсинга в первичных РНК генов человека.

II. Методика статистического подхода к предсказанию событий альтернативного сплайсинга

Идея статистического 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, объединении близкорасположенных объектов РНК в сходные кластеры. Основные этапы методики подхода с учетом выбранных наиболее оптимальных алгоритмов интеллектуального анализа данных:

Этап 1. Анализ полного набора признаков экзонов с использованием метода главных компонент [5]. Шкалирование и центрирование данных. Отбор главных компонент, объясняющих 95% вариации в данных.

Этап 2. Иерархическая кластеризация [5] экзонов гена на основе отобранного набора новых признаков (главных компонент). Разбиение экзонов на кластеры и сопоставление каждому кластеру уникального индекса в символах латинского алфавита (от a до z).

Этап 3. Преобразование символов последовательностей транскриптов РНК (от имен экзонов) к меткам кластеров, в которых расположены соответственные экзоны.

Этап 4. Удаление транскриптов дубликатов.

Этап 5. Вычисление расстояний между транскриптами РНК. Иерархическая кластеризация пула уникальных (в смысле не дубликатов) транскриптов. Представление результатов анализа в виде дендрограммы.

III. Описание вычислительного эксперимента

Для проверки работоспособности разработанных алгоритмов подхода рассмотрено предсказание событий сплйсинга на примерах экс-

периментально подтвержденных транскриптов различных пар модельных генов. В случае успешной работы подхода траскрипты РНК различных генов должны быть предсказаны с высокой точностью.

Для определения наиболее оптимального выбора метрического расстояния и метода связывания объектов в алгоритмах иерархического кластерного анализа рассмотрен кофенетический корреляционный коэффициент [5]. В качестве мер сравнения последовательностей транскриптов РНК генов исследованы различные алгоритмы лексиграфического анализа, такие как расстояния Левенштейна, Дамерау-Левенштейна, наибольшей подстроки, q-грамм, Джакарда, Джаро и Джаро-Винклера [6]. Для оценки точности предсказания принадлежности транксриптов РНК к заданному гену используется точность классификации (в %) A = 100*(N1+N2)/N, где N1 и N2 — число правильно классифицированных транскриптов для двух генов, N — общее число транскриптов двух генов.

IV. Результаты

В ходе анализа полного набора признаков экзонов с использованием метода главных компонент отобрано менее 100 наиболее значимых компонент (из более чем 1400 исходных признаков экзонов), объясняющих 95% вариации в данных. Иерархический кластерный анализ с использованием расстояния Минсковского и связывания кластеров по Варду [5] является наиболее эффективным как для группировки экзонов генов, так и для нахождения классов схожих транскриптов.

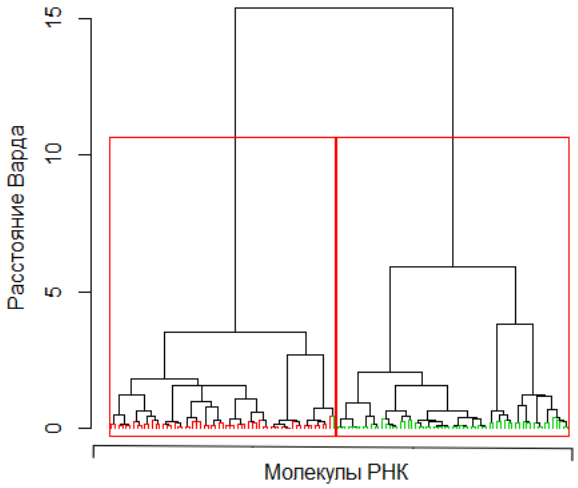

Рис. 1 – Результаты работы алгоритма иерархической кластеризации пула уникальных экспериментальных транскриптов двух модельных генов (красный и зеленый цвет)

Расстояние Джаро-Винклера с оптимизацией параметра штрафа является наилучшим, обеспечивая более высокую точность кластеризации при меньшем числе используемых кластеров. Точность разделения составляет 90-95% для рассмотренных пар модельных генов.

Таким образом, на примерах модельных генов подтверждена работоспособность разработанного подхода: экспериментально подтвержденные транскрипты пар модельных генов разделяются на два класса (рис. 1).

V. Заключение

Впервые предложен и исследован на модельных генах человека статистический подход к сравнению транскриптов генов человека, основанный на применении алгоритмов снижения размерности данных, иерархического кластерного анализа, сравнения символьных последовательностей. Точность предсказанию событий альтернативного сплайсинга состаляет более 90% для рассмотренных пар модельных генов. Предложенные алгоритмы могут быть использованы 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.

VI. Список литературы

1. Hang, J. Structural basis of pre-mRNA splicing / J. Hang, R. Wan, C. Yan, Y. Shi // Science – 2015. – Vol. 349, № 6253. – P. 1191–1198.
2. Barash, Y. Deciphering the splicing code / Y. Barash, J. A. Calarco, W. Gao, Q. Pan, X. Wang, O. Shai, B. J. Blencowe, B. J. Frey // Nature -– 2010. – Vol. 465, № 729. – P. 53–59.
3. Grinev, V. V. Decoding of exon splicing patterns in the human RUNX1-RUNX1T1 fusion gene / V. V. Grinev, A. A. Migas, A. D. Kirsanava, O. A. Mishkova, N. Siomava, T. V. Ramanouskaya, A. V. Vaitsiankova, I. M. Ilyushonak, P. V. Nazarov, L. Vallar, O. V. Aleinikova // Int. J. Biochem. Cell Biol. – 2015. – Vol. 68. – P. 48–58.
4. Aken, B. L. The Ensembl gene annotation system / B. L. Aken, S. Ayling, D. Barrell, L. Clarke, V. Curwen, S. Fairley et al. // Database (Oxford) – 2016. – Database URL: http://www.ensembl.org/index.html.
5. Интеллектуальный анализ данных / Н. Н. Яцков – Минск: БГУ, 2014. – 151 с.
6. Cohen, W. A comparison of string metrics for matching names and records / W. W. Cohen, P. Ravikumar, S. E. Fienberg // KDD Workshop on Data Cleaning and Object Consolidation – 2003. – Vol.3. – P. 73–78.