Министерство образования Республики Беларусь
Учреждение образования
«Полоцкий государственный университет»

ИНФОРМАЦИОННО-КОММУНИКАЦИОННЫЕ ТЕХНОЛОГИИ: ДОСТИЖЕНИЯ, ПРОБЛЕМЫ, ИННОВАЦИИ (ИКТ-2018)

Электронный сборник статей
I Международной научно-практической конференции,
посвященной 50-летию Полоцкого государственного университета

(Новополоцк, 14–15 июня 2018 г.)

Новополоцк
Полоцкий государственный университет
2018

Информационно-коммуникационные технологии: достижения, проблемы, инновации (ИКТ-2018) [Электронный ресурс] : электронный сборник статей I международной научно-практической конференции, посвященной 50-летию Полоцкого государственного университета, Новополоцк, 14–15 июня 2018 г. / Полоцкий государственный университет. – Новополоцк, 2018. – 1 электрон. опт. диск (CD-ROM).

Представлены результаты новейших научных исследований, в области информационно-коммуникационных и интернет-технологий, а именно: методы и технологии математического и имитационного моделирования систем; автоматизация и управление производственными процессами; программная инженерия; тестирование и верификация программ; обработка сигналов, изображений и видео; защита информации и технологии информационной безопасности; электронный маркетинг; проблемы и инновационные технологии подготовки специалистов в данной области.

Сборник включен в Государственный регистр информационного ресурса. Регистрационное свидетельство № 3201815009 от 28.03.2018.

Компьютерный дизайн М. Э. Дистанова.

Технические редакторы: Т. А. Дарьянова, О. П. Михайлова.

Компьютерная верстка Д. М. Севастьяновой.

211440, ул. Блохина, 29, г. Новополоцк, Беларусь
тел. 8 (0214) 53-21-23, e-mail: irina.psu@gmail.com

ISBN 978-985-531-602-3

УДК 575.112:004

ИССЛЕДОВАНИЕ РАКОВЫХ НОВООБРАЗОВАНИЙ МЕТОДАМИ БИОИНФОРМАТИКИ В ЭКСПЕРИМЕНТАХ ГЕНОМНОГО СЕКВЕНИРОВАНИЯ

М.К. ЧЕПЕЛЕВА
(Белорусский государственный университет, Минск);
канд. физ-мат. наук П.В. НАЗАРОВ
(Luxembourg Institute of Health, Luxembourg)

Введение. Карцинома (рак) – заболевание, характеризуемое неограниченным, неконтролируемым ростом клеток. Опухолевый процесс возникает под влиянием онкогенных факторов, которые реализуют свое действие через генетический аппарат клетки [1]. Поэтому одним из передовых направлений исследований карциномы является поиск генетических закономерностей в клетках раковых опухолей. Профилирование экспрессии с использованием анализа секвенирования РНК является мощным инструментом идентификации генов, экспрессия которых специфически изменяется в раковых клетках.

При сравнении экспрессии образцов больных и здоровых пациентов выявляются гены, которые имеют разный уровень, то есть дифференциально экспрессированны (выражены). Такие гены могут оказывать влияние на развитие раковых опухолей и подлежат дальнейшим исследованиям.

Биофункции представляют собой комплексный биологический феномен, обусловленный набором генов. Библиотеку биофункций можно визуализировать в виде направленного ациклического графа, то есть биофункция может состоять в родительских или дочерних отношениях с несколькими другими [2]. Каждая биофункция имеет список генов, которые влияют на нее. Говорят, что биофункция обогащена дифференциально выраженными генами, если список ее генов содержит статистически значимый процент таких генов. Имея список обогащенных функций, можно судить о том, в каких процессах происходят изменения при том или ином заболевании человека.

Эксперименты восстановления последовательностей ДНК и их последующий анализ проводят во многих научных лабораториях мира. Зачастую выводы достаточно сильно отличаются и не подтверждают результаты предыдущих исследований. Поэтому важно не только получить результат, но и понять, насколько он соотносится с исследованиями других научных лабораторий. Для этого необходимо исследовать адекватность результатов гентического анализа наборов экспериментальных данных, полученных из различных научных лабораторий.

Цель данного исследования – анализ экспериментальных данных геномного секвенирования, полученных из различных научных лабораторий, с использованием алгоритмов поиска дифференциально-выраженных генов и биологических функций, обогащенных данным генами; оценка устойчивости групп генов и биофункций к изменению исходной выборки, над которой проводится эксперимент.

Секвенирование РНК. Секвенирование рибонуклеиновой кислоты – технология для определения первичной структуры молекулы РНК, позволяющая получить количественную меру экспрессии генов. При помощи некоторых манипуляций матричную РНК разделяют на короткие фрагменты и восстанавливают их первичную структу-

ру. Затем создается библиотека полученных фрагментов (считываний) и специальный алгоритм отображает (картирует) считывания на эталонный геном и определяет область, где располагалось считывание. Далее выделяется набор областей, где располагаются гены или экзоны, и считается качество и количество считываний для каждой из них. Мерой экспрессии транскрипта (РНК, образовавшаяся в результате экспрессии гена) может выступать величина RPKM (Reads per kilobase per million mapped reads – число считываний на тысячу нуклеотидов на миллион картированных считываний):

$$\mathrm{RPKM} = \frac{X}{(\frac{l}{10^3}) \times (\frac{N}{10^6})}, \quad (1)$$

где X – количество считываний, попавших на транскрипт; l – длина транскрипта; N – общее количество прочтений [3].

Анализ экспрессии генов. В исследовании использовались данные экспрессии генов плоскоклеточного рака легкого TCGALUSC-dataset [4]. После фильтрации и нормализации к данным применялся алгоритм поиска дифференциально выраженных генов Limma [5].

На вход алгоритма подаются значения экспрессии генов для образцов и фактор, содержащий информацию о том, образец взят у здорового пациента или нет. Далее происходит построение матрицы, которая хранит информацию о типе каждого образца. Для каждого гена строится линейная модель:

$$E[y_j] = X \times \alpha_j, \quad (2)$$

где y_j –данные экспрессии для гена *j*; X – расчетная матрица; α_j – вектор коэффициентов или параметров линейной регрессионной модели. Коэффициенты пересчитываются два раза с учетом типа образца и применением эмпирического метода Байеса [5]. После выполняются тесты гипотез по определению класса гена. Вычисляются t-тесты для каждого гена и для каждого образца, корректируются p-значения. В результате алгоритм выдает список генов, которые дифференциально экспрессированны и характеристики проведенных тестов.

Вторая часть анализа заключается в поиске биофункций, обогащенных найденными дифференциально экспрессированными генами. Анализ выполнялся с помощью пакета topGO [6]. Входными данными являются список генов с рассчитанными на предыдущем шаге статистиками, а также база данных биофункций. Вычисляется точный критерий Фишера с учетом иерархических отношений биофункций, определяется степень «обогащения» биофункций. Поиск биофункций достаточно трудоемкий процесс и преимущество пакета topGO в том, что производится прямой подсчет генов, используется не полный набор генов, а только дифференциально экспрессированных.

В работе исследуются десять здоровых и больных пациентов. Производится поиск дифференциально выраженных генов и обогащенных биофункций для выборок из девяти здоровых и девяти больных пациентов. Используя идею метода перекрестной проверки, перебираются все возможные варианты выбора девяти из десяти здоровых и девяти из десяти больных пациентов. Для расчета похожести списков генов и списков биофункций используется индекс Жаккара:

$$J = \left\{ \frac{A \cap B}{A \cup B} \right\}. \quad (3)$$

На рисунке представлена гистограмма индексов Жаккара, рассчитанных в ходе перекрестной проверки.

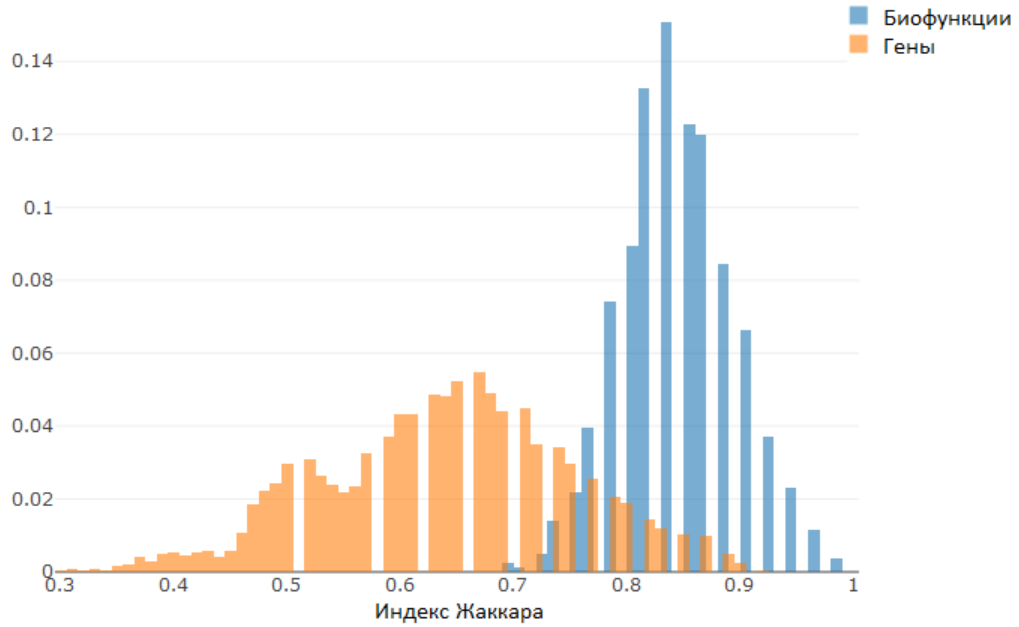

Рисунок. – Гистограмма индексов Жаккара для оценки похожести списков генов и биофункций

Списки биофункций имеют большую степень похожести, чем гены: *<J> = 0,63* для генов и *<J> = 0,84* для биологических функций. Следовательно, изменения выраженности разных генов в результате оказывает влияние на одни и те же биофункции. Замена одного пациента в исследуемой выборке вносит в среднем на 20% большую различность в гены, чем в биофункции.

Проверим, как соотносятся результаты двух независимых экспериментальных исследований (образцы взяты у разных людей). Будем формировать пары случайных выборок пациентов из доступных данных. В таблице представлены средние рассчитанные индексы Жаккара для списков генов и биофункций для различных входных выборок.

Таблица. – Индексы Жаккара в зависимости от исследуемой выборки пациентов

Размер выборки, шт. здоровые / больные	Средний индекс Жаккара для генов	Средний индекс Жаккара для биофункций
5 / 10	0,19	0,48
5 / 20	0,32	0,55
5 / 40	0,42	0,65

При исследовании малого количества образцов похожесть результатов достаточно мала, что отражает проблему соотношения результатов, к которым приводят научные исследования, проведенные в различных лабораториях.

Выводы. Выполнен анализ экспериментальных данных плоскоклеточного рака легкого, полученных в экспериментах геномного секвенирования, методами поиска дифференциально-выраженных генов и биофункций, обогащенных данными генами.

Исследована устойчивость групп генов и биофункций к изменению исходной выборки. Методом перекрестной проверки оценена похожесть списка биофункций, которая оказалась в среднем на 20% выше похожести списка генов. Установлено, что исследования, проводимые на малых выборках, составленных из разных источников экспериментальных данных, имеют низкую степень похожести для генов и невысокую для биологических функций. Это говорит о том, что малые выборки не дают достоверных результатов и необходимо продолжать интеграции научных исследований и обмен данными между лабораториями.

Литература

1. Патофизиология : учебник / под ред. В.В. Новицкого, Е.Д. Гольдберга, О.И. Уразовой. – 2009. – Т. 1.
2. Plessis, L. The what, where, how and why of gene ontology—a primer for bioinformaticians / L. Plessis, N. Škunca, C. Dessimoz // Brief Bioinform. – 2011. – Nov; 12(6). – P. 723–735.
3. Mapping and quantifying mammalian transcriptomes by RNA-Seq / A Mortazavi: Nature Methods – 2008 – 5: 621-628. PMID 18516045.
4. Genomic Data Commons [Электронный ресурс] / National Canser Institute. – Режим доступа: https://gdc.cancer.gov. – Дата доступа: 15.03.2018.

Ritchie, M.E. Limma powers differential expression analyses for RNA-sequencing and microarray studies / M.E. Ritchie, B. Phipson / Nucleic Acids Research, 43(7), 2015.

~~6.~~5. Alexa A., Rahnenfuhrer J. topGO: Enrichment Analysis for Gene Ontology. R packageversion 2.32.0, 2016.