ИССЛЕДОВАНИЕ СТАТИСТИЧЕСКОЙ УСТОЙЧИВОСТИ ДИФФЕРЕНЦИАЛЬНО ВЫРАЖЕННЫХ ГЕНОВ И ОБОГАЩЕННЫХ БИОЛОГИЧЕСКИХ ФУНКЦИЙ РАКОВЫХ КЛЕТОК

Чепелева М. К., Яцков Н. Н., Назаров П. В.
Кафедра системного анализа и компьютерного моделирования, Белорусский государственный университет
Luxembourg Institute of Health
Минск, Республика Беларусь; Luxembourg, Luxembourg
E-mail: maryna.chepeleva@gmail.com, yatskou@bsu.by, petr.nazarov@lih.lu

В работе исследуется статистическая значимость результатов анализа данных секвенирования молекул РНК плоскоклеточного рака лёгкого, полученных в разных научно-исследовательских лабораториях. Рассматриваются алгоритмы поиска дифференциально выраженных генов и обогащенных биологических функций, оценивается статистическая устойчивость групп генов и биологических функций к изменению объема экспериментальной выборки.

Введение

Рак – заболевание, характеризуемое неконтролируемым ростом клеток, возникающее под влиянием онкогенных факторов, реализующих свое действие через генетический аппарат клетки. Секвенирование РНК – технология определения первичной структуры молекулы РНК, позволяющая получить количественную меру экспрессии генов и предоставляющая экспериментальные данные для интеллектуального анализа. Гены, уровень экспрессии которых оказывается разным у тестируемых и контрольных образов, называются дифференциально выраженными и подлежат дальнейшим исследованиям [1].

Биологические функции представляют собой комплексный биологический феномен, обусловленный набором генов. Биофункция формирует иерархическую структуру с другими и включает в свою аннотацию набор генов, определяющих ее работу [2]. Биофункция обогащена дифференциально выраженными генами, если список её генов содержит значительно больше дифференциально выраженных генов, чем предсказано исходя из гипергеометрического распределения. Зная список обогащенных функций, можно точно определить, в каких процессах происходят изменения при заболевании человека.

Секвенирование РНК опухолевых тканей проводят во многих научных лабораториях мира. Часто результаты анализа достаточно сильно отличаются и не подтверждают ранее полученные выводы. Поэтому важной задачей является оценка адекватности результатов анализа наборов экспериментальных данных, полученных из разных научных лабораторий. Цель данного исследования – оценка устойчивости групп генов и биологических функций к изменению объёма выборки пациентов. Для этого был проведён анализ экспериментальных данных геномного секвенирования, полученных из различных научных лабораторий, с использованием алгоритмов поиска дифференциально выраженных генов и биологических функций, обогащенных данным генами.

I. Анализ экспрессии генов

В исследовании использовались данные экспрессии генов плоскоклеточного рака легкого TCGA LUSC-dataset [3]. После фильтрации и нормализации к данным применялся алгоритм поиска дифференциально выраженных генов на основе модифицированного критерия Стьюдента. В алгоритме производится построение матрицы, которая хранит информацию о типе каждого образца. Для каждого гена строится линейная модель:

$$E[y_j] = X\alpha_j,$$

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 [4]. Осуществляется проверка гипотез о дифференциальной выраженности генов с помощью критерия Стьюдента для каждого гена и для каждого образца, корректируются p-значения. Результатом анализа является список дифференциально выраженных генов.

Вторая часть анализа заключается в поиске биологических функций, обогащенных найденными дифференциально выраженными генами. Анализ выполнялся с использованием алгоритма на основе точного критерия Фишера [5]. Входными данными являются список дифференциально выраженных генов с рассчитанными на предыдущем шаге статистиками, а также база биологических функций. Вычисляется точный критерий Фишера с учетом иерархических отношений биофункций, определяется степень «обогащения» биофункций.

II. Исследование устойчивости

Для сравнения результатов анализа наборов данных разных размеров используются две

метрики сходства: индекс Жаккара

$$J = \frac{(A \bigcap B)}{(A \bigcup B)}$$

и косинусное расстояние – косинус угла между двумя многомерными векторами весов в пространстве биологических функций [6]. Косинусное сходство позволяет помимо пересечения учитывать степень обогащения при сравнении двух списков биологических функций.

Для оценки значимости одного пациента в малой выборке методом перекрестной проверки проводился анализ возможных вариантов выбора девяти из десяти здоровых и девяти из десяти больных пациентов. Установлено, что в среднем списки биофункций имеют большую степень сходства, чем гены: $< J >= 0,63$ для генов и $< J >= 0,84$ для биологических функций. Изменения выраженности разных генов в результате могут оказывать влияние на одни и те же биологические функции. Замена пациента в исследуемой выборке вносит в среднем на 20% большую информативность в гены, чем в биофункции.

В случае сравнения двух независимых экспериментальных исследований (образцы взяты у разных людей) были сформированы пары случайных выборок пациентов из доступных данных. В таблице 1 представлены средние рассчитанные индексы Жаккара для списков генов и биологических функций для наборов данных разного размера.

Таблица 1 – Индексы Жаккара в зависимости от объёма исследуемой выборки

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
10 / 40	0,42	0,61
10 / 80	0,54	0,66
10 / 100	0,58	0,69
25 / 250	0,76	0,81

При рассмотрении малого количества образцов достоверность результатов анализа достаточно мала, что отражает проблему соотношения результатов независимых исследований. Увеличение размеров выборки повышает идентичность результатов, что свидетельствует о необходимости обмена данными между научными лабораториями.

Необходимо оценить, насколько одно отдельно взятое исследование небольшого количества пациентов будет соотноситься с результатами анализа большого «эталонного» набора данных. Исследовался набор из 550 образцов (50 здоровых и 500 больных). Результаты сравнительного анализа представлены в таблице 2.

Таблица 2 – Индексы Жаккара и косинусные сходства результатов в зависимости от объёма исследуемой выборки

Размер малой выборки, шт. здоровые / больные	Средний индекс Жаккара для генов	Средний индекс Жаккара для биофункций	Среднее косинусное сходство для биофункций
5 / 10	0,23	0,19	0,20
5 / 20	0,30	0,25	0,29
10 / 40	0,51	0,57	0,82
10 / 100	0, 61	0,68	0,89
25 / 250	0,75	0,84	0,98

Сходство результатов на малых выборках с результатом большой выше, чем при сравнении экспериментов на малых с малыми. Косинусное сходство выше, чем индекс Жаккара для биологических функций, что подтверждает то, что биологические процессы, наиболее сильно подвергающиеся изменениям при болезни, в большей степени схожи, чем те, которые испытывают менее сильное влияние.

Заключение

Представленный анализ экспериментальных данных секвенировая РНК позволяет определить дифференциально выраженные гены и биофункции, обогащенные данными генами. Результаты исследований, проводимых на малых выборках, составленных из разных источников экспериментальных данных, имеют низкую степень идентичности. При повышении размера выборки подобие возрастает, что подтверждает необходимость продолжать интеграцию научных исследований и экспериментальных данных геномного секвенирования. Устойчивость списков дифференциально выраженных генов ниже устойчивости биофункций. Для получения статистически значимых результатов достаточно рассмотреть 100 и более раковых образцов.

1. Nazarov, P. V. RNA sequencing and transcriptome arrays analyses show opposing results for alternative splicing in patient derived samples. / P. V. Nazarov [et al.] // BMC Genomics – 2017. – Vol. 18. – Ar. num. 443.
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. – Vol. 12. – P. 723–735.
3. Genomic Data Commons [Electronic resource] : National Cancer Institute. – Mode of access: https://gdc.cancer.gov. – Date of access: 15.09.2018.
4. Ritchie, M. E. Limma powers differential expression analyses for RNA-sequencing and microarray studies / M. E. Ritchie, B. Phipson // Nucleic Acids Research. – 2015. – Vol. 43. – P. e47.
5. Alexa, A. Enrichment Analysis for Gene Ontology with topGO [Electronic resource] / A. Alexa, J. Rahnenfuhrer // Bioconductor. – Mode of access: https://bioconductor.org/packages/release/bioc/html/topGO.html – Date of access: 4.09.2018.
6. Manning, C. D. An Introduction to Information Retrieval Draft / C. D. Manning, P. Raghavan, H. Schütze // Online edition. Cambridge University Press. – 2009. – P. 121.