УДК 57.087.1

**Исследование чувствительности метода анализа однородности экспрессионных данных**

**Алиев Р.О.\*1, Борисов Н.М\*\*2.**

1*НИЦ "Курчатовский институт", Москва, Россия* 2*Первый Московский государственный медицинский университет им. И.М.Сеченова, Москва, Россия*

***Аннотация.*** Развитие биоинформатики делает вопрос однородности депонированных в открытых источниках данных, полученных в результате профилирований уровней экспрессии генов, всё более актуальным из-за роста требований, предъявляемых к входным данным.

Ранее нами был предложен метод анализа однородности данных генной экспрессии, обсуждены первые результаты работы метода на реальных и искусственно сгенерированных данных, показаны его некоторые преимущества. Суть метода заключалась в последовательных сравнениях случайных разбиений данных при помощи теста Стьюдента. Результатом работы метода является график плотности корреляций величин, зависящих от значения p-value и средних значений экспрессии генов, по которому можно судить об однородности данных. В настоящей статье рассказано о проведённых нами вычислительных экспериментах на модельных данных, расширяющих возможности интерпретации работы авторского метода, что позволит исследователю делать более полные выводы о тех данных, с которыми он работает.

***Ключевые слова:*** генная экспрессия, профилирование мРНК, транскриптом, большие данные, репозитории общедоступных данных, тест Стьюдента, кластеризация.

ВВЕДЕНИЕ

На сегодняшний день в базе данных открытого доступа GEO содержится 44477 серий, содержащих более 1,5 млн. профилей генной экспрессии человеческих образцов [1], полученных при помощи микрочиповых исследований и методов секвенирования нового поколения. Как правило, для таких серий, даже полученных на одной и той же экспериментальной платформе и относящихся к одной и той же нозологической форме, в настоящее время нет однозначного ответа, являются ли данные о генной экспрессии, однородными.

В предыдущей статье мы предложили новый метод анализа однородности экспрессионных данных на основе теста Стьюдента. Он был использован на реальных и сгенерированных в ходе модельных расчетов данных, сравнён с методом иерархической кластеризации в чувствительности и быстродействии, были продемонстрированы дополнительные возможности метода [2]. Нами были исследованы материалы (серии экспериментов), депонированные в базе данных открытого доступа GEO [3]. Они содержали в себе информацию об уровнях генной экспрессии различных здоровых и патологических тканей, таких как: серое и чёрное вещество, микроглия, мышечная и почечная ткань, кровь и костный мозг. Патологические образцы были получены у пациентов, страдающих от болезни Альцгеймера, Паркинсона, венозной тромбоэмболии, глиобластомы и нейробластомы [4-17].

Однако, обработанных реальных данных было недостаточно для полного понимания их влияния на конечный результат. В настоящей работе мы оценили влияние на результат особенностей распределения исходных данных, исследовав сгенерированные с различными видами и параметрами распределений. Обработка таблиц с последовательно добавленной неоднородностью в исследуемые данные позволила уточнить чувствительность метода, а относительно невысокая времязатратность при использовании большого количества данных в рамках одного вычислительного эксперимента говорит о возможности реализации автоматической системы мониторинга однородности депонированных данных в базах открытого доступа.

**МЕТОДЫ**

**Оценка однородности экспрессионных профилей**

Запишем исследуемые экспрессионные данные в виде таблиц, по строкам которых записаны значения уровней экспрессии генов, а в столбцах - образцы. Образцы могут быть получены как из здоровых тканей, так и из патологических. Для удобства перед обработкой все здоровые образцы (H - healthy) перемещаются в левую часть таблицы, патологические (D - desease) - в правую.

Нами [2] был предложен метод анализа однородности экспрессионных данных. Во время его работы исследуемые профили патологических образцов случайно разбиваются на две группы (при этом от профили здоровых образцов не разбиваются), после чего каждая группа сравнивается по генам с уровнем экспрессии здоровых тканей при помощи формулы, использующей тест Стьюдента

, (1)

после чего к полученным значениям для генов в каждой группе данных рассчитывается Пирсоновский коэффициент корреляции. Данный процесс повторяется для новых случайных разбиений. Строится график плотности полученных коэффициентов корреляции, по свойствам которого (его симметричность, положение и количество мод) исследователь сможет сделать выводы об уровне однородности данных. Подробное описание работы метода было дано в предыдущей статье [2], иллюстрация приведена на рисунке 1.



а)



б)

**Рис. 1.** Иллюстрация используемого метода: **а)** разбиение исходных данных, содержащих результаты транскриптомных экспериментов, на две таблицы и расчёт величины *f* для каждого гена (см. формулу (1)); **б)** расчёт плотности коэффициентов корреляции для столбцов *f*, полученных для всех созданных разбиений.

**Модельные вычислительные эксперименты**

В качестве материалов для исследования были использованы сгенерированные в ходе вычислительных экспериментов модельные данные. Законы и параметры распределения уровней экспрессий выбирались с целью исследовать возможности нашего метода анализа неоднородности экспрессионных данных. Нашей целью было проверить способности предложенного метода различать как распределения одинакового типа с разными параметрами, так и распределения разных типов. Одним типом распределения служило нормальное, математическое ожидание и стандартное отклонение которого были индивидуальны для каждого гена; другим, который мы хотели отличить от первого с помощью нашего метода, служило равномерное распределение *U*, принимавшее значения от 0 до max(Ni). Количество генов для генерируемых таблиц составляло 10000 для первого вычислительного эксперимента и 5000 для всех последующих, в зависимости от объёма необходимых вычислений. Количество образцов, считающихся здоровыми, составляло 20, считающихся патологическими - 40.

Для исследования зависимости результатов работы метода от параметров, характеризующих распределение величин экспрессий генов, был проведён вычислительный эксперимент №1. Было сгенерировано 13 различных вариантов входных данных, в каждом по 10000 генов. Структуру распределений входных данных см. на рис. 2.1-13. Расчёт был проведён для четырёх наборов параметров нормальных распределений однородных данных *N* (см. таблицу 1). В первых двух случаях математическое ожидание нормального распределения равномерно принимало значения *μi* ϵ[1;300], в двух других - *μi* ϵ[1;700]. Стандартное отклонение *σi* равномерно выбиралось на интервале (0;300] для одной пары случаев на интервале (0;10] для другой; при достижении отрицательных значений всё распределение смещалось вправо так, чтобы ноль быль крайним левым значением (см. таблицу 1).

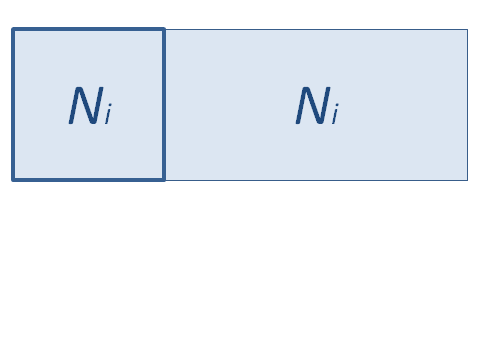
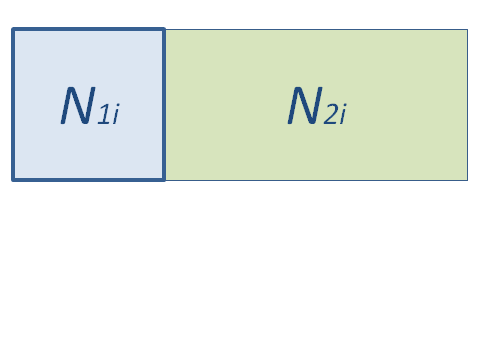
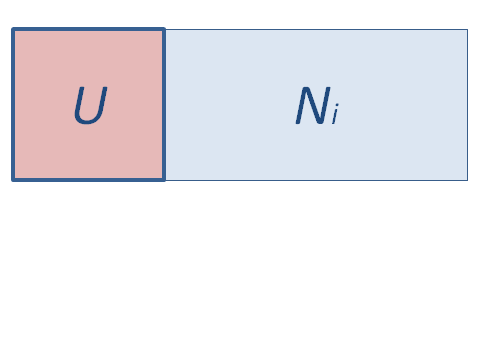
**Таблица 1.** Диапазон принимаемых значений параметров нормальных распределений для первого вычислительного эксперимента

|  |  |  |
| --- | --- | --- |
| Номер расчёта | Среднее значение | Стандартное отклонение |
| 1 | [1;300] | (0;300] |
| 2 | [1;300] | (0;10] |
| 3 | [1;700] | (0;300] |
| 4 | [1;700] | (0;10] |

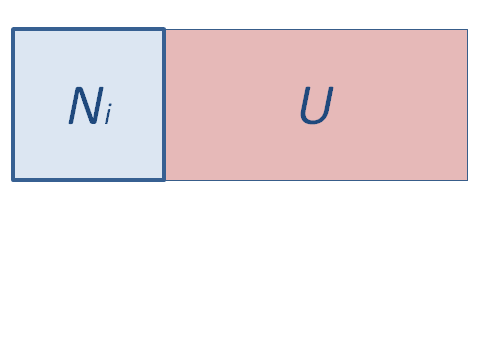
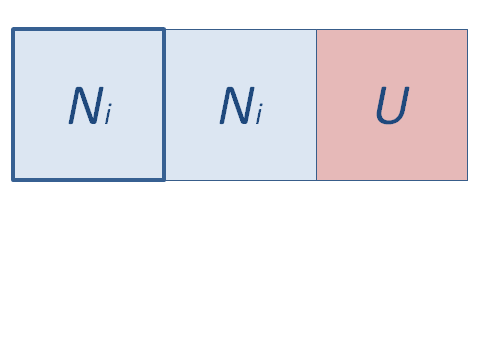
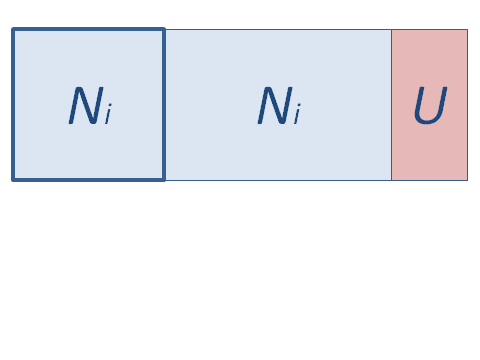
Для определения факторов, влияющих на положение пика в результирующем графике нами был поставлен вычислительный эксперимент №2 на сгенерированных природоподобных данных. Отрицательно-биномиальное распределение в достаточной степени хорошо описывает реальный характер распределения экспрессии генов [18]. Нами было принято решение продолжить использование нормального распределения, так как оно является предельным случаем отрицательно-биномиального. Генерация данных была усложнена, величина *μi* выбиралась из равномерного распределения на отрезке [1;10000], а *σi* – из равномерного распределения на отрезке от 10% до 35% от *μi*. Таким образом была проведена попытка генерации более природоподобных данных, когда уровень экспрессии может отличаться в разы от гена к гену, при этом с разным стандартным отклонением для каждой строки. Таблица значений экспрессий генов схематически изображена на рис 2.1. Столбцы таблицы соответствовали отдельным экспрессионным профилям, строки – генам; общее количество генов было равно 5000. Уровни экспрессии для каждого i-ого гена в здоровых и патологических образцах подчинялась одному и тому же распределению (с одинаковыми параметрами *μi* и *σi*). Для обеспечения различия здоровых и патологических образцов среднее значение и стандартное отклонение нормального распределения, из которого генерировались случайные значения уровней экспрессий генов в патологических образцах, умножались на корректирующие коэффициенты. Таким образом, в каждой сгенерированной таблице нормальное распределение *i*-ого гена для патологических образцов имело вид *Ni* (*μik,* (*σiz*)2), где *k* ϵ{0.2, 0.25, 0.3, ..., 2.5} (всего 47 значений), а *z* ϵ{0.25, 0.5, 0.75, ..., 2} (всего 8 значений). Общее количество сгенерированных таблиц составило 376.

Для определения влияния построчного добавления неоднородности в данные на результирующий график был проведён вычислительный эксперимент №3 (рис. 2.15). Исходные данные были такими же, как и для второго (природоподобная генерация, структура таблицы как на рис.2.1). Затем в правой части таблицы строки *Ni* последовательно заменялись на равномерное распределение *U* с шагом в 1% от общего числа генов, всего было сгенерировано 100 таблиц

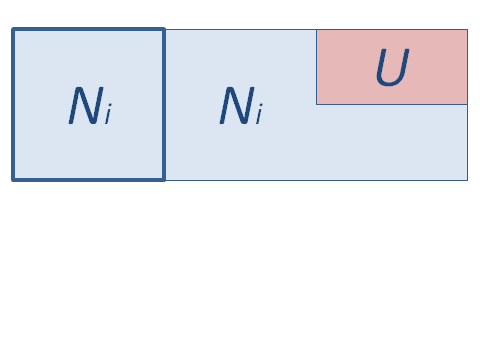
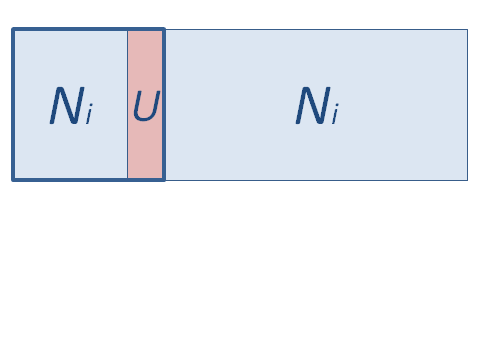
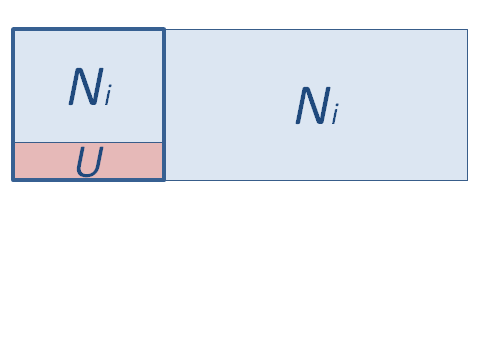
Для исследования чувствительности авторского метода и метода иерархической кластеризации при помощи последовательного добавления неоднородности по столбцам и строкам входных данных были проведены вычислительные эксперименты №№ 4 и 5 (рис. 2.16-17). Они являются усложнённым вариантом третьего. Для них тоже использовалась природоподобная генерация *Ni*. Исходный вид таблицы для четвёртого эксперимента был как на рис.2.2, для пятого - как на рис.2.1. В качестве замены использовались природоподобные данные, подчиняющиеся нормальному распределению с другими параметрами в каждой строке, сама замена происходила не целыми строками, как в эксперименте №3, а последовательно увеличивающимися частями строк, так, в начале замена была в одном образце, потом в нескольких, в конце во всех (целой строкой). Для четвёртого эксперимента пошаговые замены проходили с 1-ого образца с шагом 1 до 21, потом первые 25, 30,35 и 40 образцов. По строкам заменялся 1 ген, потом 5, 25, 100, 250, потом с шагом 250 генов до 5000, всего было сгенерировано 626 таблиц. Для пятого эксперимента было решено использовать шаг, равный 5, для столбцов, правила для строк (генов) остались такими же, как и в четвёртом, всего сгенерировано 226 таблиц.

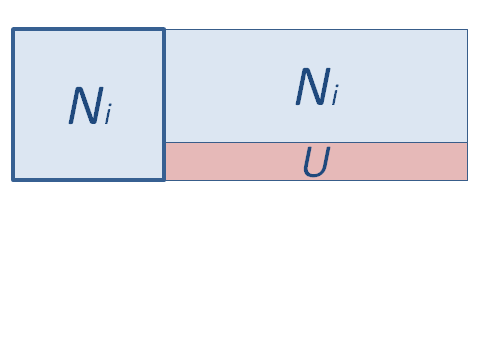
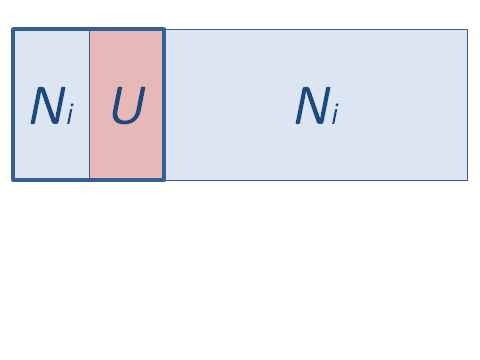
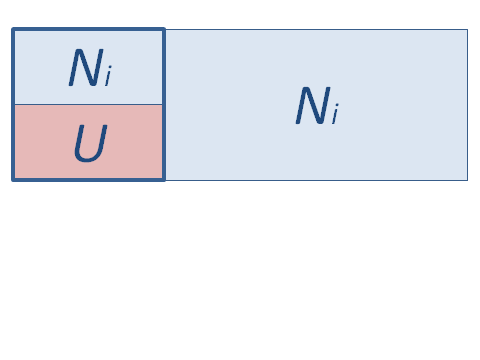
1) 2) 3)

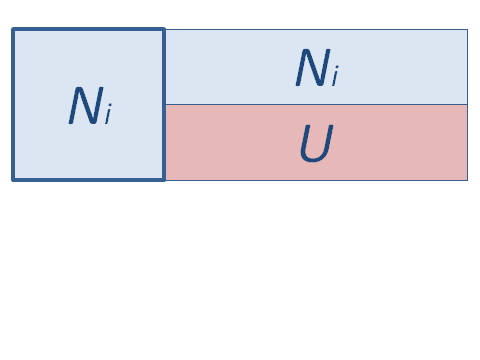
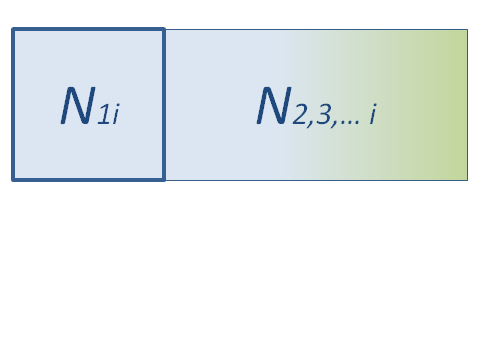
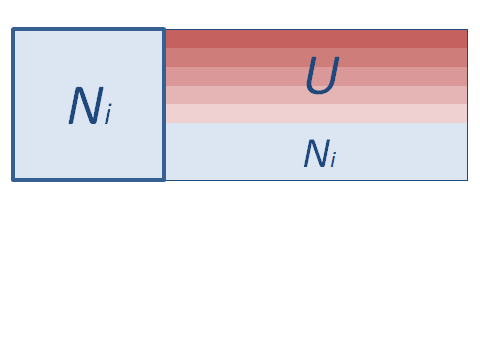
4) 5) 6)

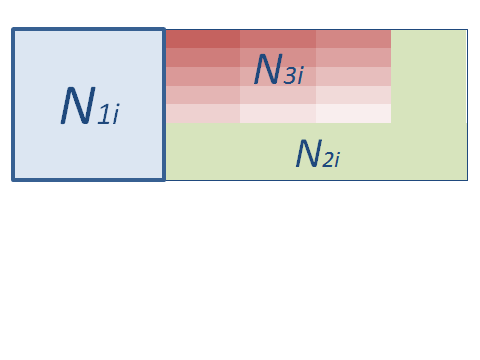
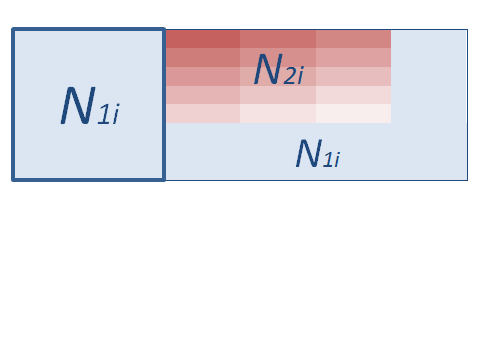
7) 8) 9)

10) 11) 12)

13) 14) 15)

16) 17)

**Рис. 2.** Иллюстрация структуры распределений в входных данных для экспериментов. 1:13) Вычислительный эксперимент №1; 14) Вычислительный эксперимент №2; 15) Вычислительный эксперимент №3; 16) Вычислительный эксперимент №4; 17) Вычислительный эксперимент №5. Рамкой в левой части всех рисунков обведены данные, соответствующие образцам здоровых тканей. Синим и зелёным выделены однородные данные с нормальными распределением, красным – инородные по отношению к нормально распределенным данным. Инородность добавляется последовательно, более светлый оттенок означает последующую итерацию.

**Используемое программное обеспечение**

Для реализации метода была использована бесплатно распространяемая среда с открытым кодом – R. В работе метод не требует дополнительно установленных библиотек, наличие базовой является достаточным (без параллельных вычислений).

В последнем эксперименте авторский метод был сравнён с методом иерархической кластеризации, для реализации которого были использованы пакеты pvclust [19] и magrittr [20]. В рамках увеличения производительности метода, было решено использовать возможности параллельно исчисления, которые были получены при помощи пакетов foreach [21] и doparallel [22].

**РЕЗУЛЬТАТЫ**

Нами было последовательно проведено пять вычислительных экспериментов, каждый из которых позволял уточнять условия и параметры для следующего. Главной нашей задачей было улучшить понимание работы метода, в том числе определить влияние структуры входных данных на результат и чувствительность к неоднородности.

**Вычислительный эксперимент №1 – Определение влияния структуры входных данных на результирующий график**

Нами была обнаружена качественная симметричность результатов обработки данных, для которых и здоровые образцы, и паталогические (на самом деле важно то, что в данном эксперименте норма менялась вместе с болезнью) одинаково меняли предел для среднего в генерируемых нормальных распределениях, то есть, для данного эксперимента влияние оказывала только величина дисперсии (рис. 3.6(а, б)). Приведённые в работе иллюстрации результатов для первого эксперимента соответствуют обработке сгенерированных данных со следующими параметрами распределений: рис.3.1-3.6(а), 3.7-3.13 - красный (μi ϵ[1;300], *σi* ϵ(0;10]), синий (μi ϵ[1;700], *σi* ϵ(0;300]); рис.3.6(б) - красный (μi ϵ[1;700], *σi* ϵ(0;10]), синий (μi ϵ[1;300], *σi* ϵ(0;300]).



1) 2) 3)



4) 5) 6.а)



6.б) 7) 8)



9) 10) 11)



12) 13)

**Рис. 3.** Результаты работы метода в первом расчётном эксперименте для всех 13-и случаев входных данных. Номер рисунка соответствует номеру структуры данных, см. рис. 2. 1-6(а),7-13: красный - μi ϵ[1;300], *σi* ϵ(0;10]; синий - μi ϵ[1;700], *σi* ϵ(0;300]. 6(б): красный - μi ϵ[1;700], *σi* ϵ(0;10]; синий - μi ϵ[1;300], *σi* ϵ(0;300].

Анализ 1-4 структур данных позволил прийти к следующим выводам: чем сильнее похожи здоровые образцы на патологические, тем левее будет находиться результирующий график, особенно хорошо это видно на рис.3.4, в случае, когда увеличение дисперсии в генерации здоровых тканей приводит к большей её схожести на равномерные данные; увеличение дисперсии в целом по всем данным приводит к смещению результирующего графика влево, что говорит о том, что два набора данных с высоким уровнем дисперсии труднее отличить. Пятый случай, когда здоровые образцы и половина образцов патологических идентичны, а вторая половина является инородной, в данном случае это равномерное распределение, показал недостаток метода, при интерпретации результата данные могли быть приняты за однородные. По этой причине было решено провести более подробное исследование данного результата, что привело к проведению четвёртого и пятого вычислительного эксперимента.

Графики 3.6(а, б) и 3.7 показывают появление в данных случаях многомодальности в результирующем графике, что позволило бы отнести такие данные к неоднородным. Случаи 8 и 11, в которых для здоровых образцов добавлена неоднородность по столбцам (четверть образцов и половина соответственно), продемонстрировали похожие результаты. Случаи 9 и 12, в которых аналогичным образом добавлена неоднородность по строкам, показали результат, идентичный второму случаю, это говорит о том, что добавление подобной неоднородности делает здоровые образцы лучше отличимыми от патологических, в отличии от неоднородности по столбцам, т.е. образцам. Так как метод используется для анализа однородности патологических данных, неоднородность здоровых образцов не приводит к появлению многомодальности в результирующем графике.

Последние случаи, а именно 10 и 13, являются, по сути, примерами однородных данных, часть которых имеет высокий уровень дисперсии, являясь равномерным распределением. В качестве результатов были получены одномодальные графики, положение которых зависело от размытости начальных данных. Данные с низкой дисперсией лучше отличимы на фоне размытых данных, чем данные с высоким уровнем дисперсии, это привело к смещению центров пиков влево на рис.3.13 относительно рис.3.10. Для более подробного анализа этих результатов было решено провести дополнительные вычислительные эксперименты.

**Вычислительный эксперимент №2 – определение факторов, влияющих на положение пика в результирующем графике**

Для подобного вида данных результатом применения метода был одномодальный график, поэтому для всех 376 таблиц записывалось среднее значение корреляции векторов *f* (что соответствует центру пика в результирующем графике). Иллюстрация результатов данного эксперимента приведена на рис. 4. Из графика можно увидеть, что увеличение дисперсии в исследуемых данных (экспрессия генов в патологических образцах) приводит к смещению результирующего пика влево (синий цвет), а изменение среднего в нормальных распределениях симметрично к увеличению и уменьшению и приводит к смещению результирующего пика вправо (красный цвет). Суммируя результаты можно сказать, чем сильнее болезнь отличима от здорового состояния, тем правее будет находиться результирующий пик для исследуемых однородных данных.



**Рис. 4.** Результаты вычислительного эксперимента №2. По оси X: относительное изменение величины стандартного отклонения для распределений экспрессии генов исследуемых образцов. По оси Y: относительное изменение центра нормальных распределений. Цвет соответствует полученным значениям корреляции (*r)* векторов *f.*

**Вычислительный эксперимент № 3 – Влияние построчного добавления неоднородности в данные на результирующий график**

Третий вычислительный эксперимент заключался в последовательной замене однородных данных на неоднородные, изменения вносились только в данные, соответствующие экспрессии патологических образцов. Так же, как и в предыдущем эксперименте результатом обработки всех сгенерированных таблиц являлся одномодальный график, поэтому для дальнейшего анализа было использовано среднее значение корреляции *f*. Результат представлен на рис. 5.



**Рис. 5.** Результаты вычислительного эксперимента №3. По оси X: процент генов (p), экспрессия которых заменена на равномерное распределение (добавленная инородность). По оси Y: положение центра результирующего пика.

Можно обратить внимание, что первое добавление неоднородности приводит к резкому смещению результирующего пика вправо. Это объясняется тем, что здоровые образцы становится более отличимы от патологических. Однако этот процесс замедляется и в конечном итоге, начиная с 36% заменённых генов, происходит смещение результирующего пика влево. Это можно объяснить тем, что полученные данные в целом начинают обладать всё большим уровнем дисперсии, приводящем к тому, что патологические образцы становится труднее отличить от здоровых.

**Вычислительные эксперименты №№ 4 и 5 – Исследование чувствительности авторского метода и метода иерархической кластеризации при помощи последовательного добавления неоднородности по столбцам и строкам входных данных**

Целью данных расчетов являлось определить чувствительность авторского метода, а также сравнить её с возможностями метода иерархической кластеризации. Для четвёртого расчёта здоровые образцы и патологические имели одинаковые параметры генерации распределения экспрессии генов, для пятого - разные. Было определено, что разница между здоровыми и больными образцами оказывает влияние на чувствительность авторского метода. Так, неоднородность данных, имеющих отличия от здоровых образцов, выявляется гораздо раньше, чем у изначально схожих данных (рис.6(а,б)).

Для случая изначально различных здоровых и патологических образцов можно наблюдать падение чувствительности при приближении к случаю внесения одновременных изменений в половину образцов. Это связано с процедурой проведения вычислений по нашему методу. Для второго случая (изначально идентичные данные) была выявлена область измененных значений уровней экспрессии генов в инородных образцах, для которой эти значения для всех патологических образцов были классифицированы как однородные. Это связано с тем, что данные, на которые происходит замена, становятся превалирующими, и уже не они становятся шумом, а изначальные данные. Эта область не симметрична той, что находится в приближении к нулю по количеству единовременно изменяемых образцов, потому что данные в этих точках отличаются по структуре. При этом предельный случай, когда все гены заменены в половине образцов, был выявлен ранее, см. рис.3.5. Для обоих случаев выявлена невозможность точно определить неоднородность данных в случае единичного аномального образца. Для борьбы с этими эффектами в метод необходимо внести небольшие изменения, например, искусственное прореживание данных, либо замена равновеликого разбиения патологических образцов на неравновеликую, 2 к 1 или 3 к 2, что позволит избежать тех случаев, когда перестановка образцов ни на что не влияет.

В отличие от нашего метода, метод иерархической кластеризации показывает стабильную пороговую чувствительность, однако были выявлено принятие ложных решений для случая, когда изменениям подвергались все 40 образцов. Также стоит отметить, что чувствительность иерархической кластеризации, начиная с 5 образцов, составляла 100 генов. При этом наш метод, начиная с 7 образцов, показывал более высокую чувствительность (от 100, до 25 инородных генов).

Стоит отметить, что для всех трёх описанных случаев, в том числе и для метода иерархической кластеризации, существует область заменённых генов в инородных образцах, для которой классификация данных затруднена. Это область перехода, в которой с добавлением неоднородности классификация результирующих данных может чередоваться (однородные/неоднородные).



а) б) в)

**Рис. 6.** Результаты вычислительных экспериментов №№ 4 и 5. Красный: данные классифицированы как неоднородные (появление многомодальности); Синий: классификация затруднена из-за неопределенности. Зелёный: данные признаны однородными. а) Авторский метод, здоровые и патологические образцы изначально различны; б) Авторский метод, здоровые и патологические образцы изначально идентичны; в) Метод иерархической кластеризации

**ЗАКЛЮЧЕНИЕ**

В статье была продемонстрирована возможность достижения высокого уровня чувствительности метода. Расчетные эксперименты позволили уточнить процесс интерпретации результатов, а объём таблиц и затраченное время на их обработку говорит о возможности использования нашего метода как для экспресс-анализа при исследовании новых данных, так и для анализа накопленного материала в базах открытого доступа.

Мы ожидаем, что авторский метод применим как для анализа микрочиповых исследований, так и для анализа результатов, полученных при помощи методов секвенирования нового поколения. Были проведены вычислительные эксперименты на реальных и искусственно сгенерированных данных, проведено сравнение авторского метода с методом иерархической кластеризации [2]. Выигрыш по времени при обработке больших массивов данных, определение наличия неоднородности с возможностью качественной её оценки, простота использования, дополнительная визуализация отклонений экспрессии генов [2] - всё это позволяет сделать авторский метод полезным для анализа больших массивов профилей экспрессии генов и генных продуктов.

**ПИСОК ЛИТЕРАТУРЫ**

1. *Gene Expression Omnibus*. URL: https://www.ncbi.nlm.nih.gov/geo/ (дата обращения: 21.01.2019).
2. Алиев Р. О., Борисов Н. М. Метод анализа однородности экспрессионных данных на основе теста Стьюдента. *Математическая биология и биоинформатика*. 2018. Т. 13. №. 1. Стр. 50-67.
3. Edgar R., Domrachev M., Lash A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research.* 2002. V. 30. № 1. P. 207–210.
4. Welle S., Brooks A.I., Delehanty J.M., Needler N., Thornton C.A. Gene expression profile of aging in human muscle. *Physiological Genomics*. 2003. V. 14. № 2. P. 149–159.
5. Blalock E.M., Geddes J.W., Chen K.C., Porter N.M., Markesbery W.R., Landfield P.W. Incipient Alzheimer's disease: microarray correlation analyses reveal major transcriptional and tumor suppressor responses. *PNAS.* 2004. V. 101. № 7. P. 2173–2178.
6. Borovecki F., Lovrecic L., Zhou J., Jeong H., Then F., Rosas H.D., Hersch S.M., Hogarth P., Bouzou B., Jensen R.V., Krainc D. Genome-wide expression profiling of human blood reveals biomarkers for Huntington's disease. *PNAS*. 2005. V. 102. № 31. P. 11023–11028.
7. Sternberg A, Killick S., Littlewood T., Hatton C., Peniket A., Seidl T., Soneji S., Leach J., Bowen D., Chapman C. et al. Evidence for reduced B-cell progenitors in early (low-risk) myelodysplastic syndrome. *Blood*. 2005. V. 106. № 9. P. 2982–2991.
8. Scherzer C.R., Eklund A.C., Morse L.J., Liao Z., Locascio J.J., Fefer D., Schwarzschild M.A., Schlossmacher M.G., Hauser M.A., Vance J.M., Sudarsky L.R. et al. Molecular markers of early Parkinson's disease based on gene expression in blood. *PNAS*. 2007. V. 104. № 3. P. 955–960.
9. Yusenko M.V., Zubakov D., Kovacs G. Gene expression profiling of chromophobe renal cell carcinomas and renal oncocytomas by Affymetrix GeneChip using pooled and individual tumours. *International Journal of Biological Sciences*. 2009. V. 5. № 6. P. 517.
10. Duke D.C., Moran L.B., Pearce R.K.B., Graeber M.B. The medial and lateral substantia nigra in Parkinson’s disease: mRNA profiles associated with higher brain tissue vulnerability. *Neurogenetics*. 2007. V. 8. № 2. P. 83–94.
11. Kaizer E.C., Glaser C.L., Chaussabel D., Banchereau J., Pascual V., White. P.C. Gene expression in peripheral blood mononuclear cells from children with diabetes. *The Journal of Clinical Endocrinology & Metabolism*. 2007. V. 92. № 9. P. 3705–3711.
12. Hokama M., Oka S., Leon J., Ninomiya T., Honda H., Sasaki K., Iwaki T., Ohara T., Sasaki T., LaFerla F.M. et al. Altered expression of diabetes-related genes in Alzheimer's disease brains: the Hisayama study. *Cerebral Cortex*. 2013. V. 24. № 9. P. 2476–2488.
13. Lewis D.A., Stashenko G.J., Akay O.M., Price L.I., Owzar K., Ginsburg G.S., Chi J., Ortel T.L. Whole blood gene expression analyses in patients with single versus recurrent venous thromboembolism. *Thrombosis Research*. 2011. V. 128. № 6. P. 536–540.
14. Lewis D.A., Suchindran S., Beckman M.G., Hooper W.C., Grant A.M., Heit J.A., Manco-Johnson M., Moll S., Philipp C.S., Kenney K. et al. Whole blood gene expression profiles distinguish clinical phenotypes of venous thromboembolism. *Thrombosis Research*. 2015. V. 135. № 4. P. 659–665.
15. Tso C.L., Shintaku P., Chen J., Liu Q., Liu J., Chen Z., Yoshimoto K., Mischel P.S., Cloughesy T.F., Liau L.M., Nelson S.F. Primary glioblastomas express mesenchymal stem-like properties. *Molecular Cancer Research*. 2006. V. 4. № 9. P. 607–619.
16. Asgharzadeh S., Pique-Regi R., Sposto R., Wang H., Yang Y., Shimada H., Matthay K., Buckley J., Ortega A., Seeger R.C. Prognostic significance of gene expression profiles of metastatic neuroblastomas lacking MYCN gene amplification. *Journal of the National Cancer Institute*. 2006. V. 98. № 17. P. 1193–1203.
17. Rock R.B., Hu S., Deshpande A., Munir S., May B.J., Baker C.A., Peterson P.K., Kapur V. Transcriptional response of human microglial cells to interferon-[gamma]. *Genes and Immunity*. 2005. V. 6. № 8. P. 712.
18. Di Y. et al. The NBP negative binomial model for assessing differential gene expression from RNA-Seq //Statistical Applications in Genetics and Molecular Biology. – 2011. – Т. 10. – №. 1.
19. Suzuki R., Shimodaira H. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics*. 2006. V. 22. № 12. P. 1540–1542.
20. Bache S.M., Wickham H. Magrittr: A forward-pipe operator for R. *R package version 1.5*. 2014. URL: https://CRAN.R-project.org/package=magrittr (дата обращения: 03.12.2018).
21. Calaway R., Weston S. foreach: Provides Foreach Looping Construct for R. *R package version 1.4.4.* 2017. https://CRAN.R-project.org/package=foreach (дата обращения: 03.12.2018).
22. Calaway R, Tenenbaum D. doParallel: Foreach Parallel Adaptor for the 'parallel' Package. *R package version 1.0.14*. 2018. https://CRAN.R-project.org/package=doParallel (дата обращения: 03.12.2018).