МОЛЕКУЛЯРНАЯ БИОЛОГИЯ, 2025, том 59, № 1, с. 154–161

БИОИНФОРМАТИКА

УДК 57.087.1

ТРАНСКРИПТОМИКА И “ПРОКЛЯТИЕ РАЗМЕРНОСТИ”: МОНТЕ-КАРЛО СИМУЛЯЦИИ КЛАССИФИКАЦИОННЫХ МОДЕЛЕЙ КАК ИНСТРУМЕНТ АНАЛИЗА МНОГОМЕРНЫХ ДАННЫХ В ЗАДАЧАХ ПОИСКА МАРКЕРОВ БИОЛОГИЧЕСКИХ ПРОЦЕССОВ

© 2025 г. Г. Ж. Осьмак*, М. В. Писклова

Национальный медицинский исследовательский центр кардиологии им. ак. Е.И. Чазова Министерства здравоохранения Российской Федерации, Москва, 121552 Россия
Российский национальный исследовательский медицинский университет им. Н.И. Пирогова Министерства здравоохранения Российской Федерации, Москва, 117997 Россия
**e-mail: german.osmak@gmail.com*

Поступила в редакцию 11.04.2024 г.
После доработки 06.05.2024 г.
Принята к публикации 26.05.2024 г.

Высокопроизводительные методы исследования транскриптома позволяют оценить огромное количество факторов, что ценно для ученых, но порождает проблему “проклятия размерности”, что повышает требования к методам обработки и анализа данных. В представленной работе мы предлагаем новый алгоритм, объединяющий методы Монте-Карло и машинное обучение. Этот алгоритм позволит сократить пространство признаков, подсвечивая гены, с наибольшей вероятностью ассоциированные с определенными заболеваниями. Представленный подход позволяет не только сформировать набор “интересных” генов, но и взвесить их множество, присвоив каждому гену меру его “важности”. Эта мера может быть использована как в последующем статистическом анализе, так и при визуализации и интерпретации результатов. Работа алгоритма продемонстрирована нами на открытых данных профилирования больных гипертрофической кардиомиопатией. По результатам анализа выявлены гены *MYH6*, *FCN3*, *RASD1* и *SERPINA3,* что хорошо согласуется с опубликованными данными.

Ключевые слова: транкриптомика, машинное обучение, Монте-Карло, гипертрофическая кардиомиопатия, биомаркеры

DOI: 10.31857/S0026898425010117, **EDN:** HCCMTU

ВВЕДЕНИЕ

Транскриптомика и высокопроизводительные методы исследования, такие как РНК-секвенирование (RNA-seq) или микрочипы (MicroArray), сегодня, без сомнений, занимают важное место в арсенале инструментов для изучения молекулярных механизмов биологических систем, патогенеза различных заболеваний и поиска их маркеров [1].

Идентификация дифференциально экспрессируемых генов (ДЭГ) или транскриптов в различных условиях (группах сравнения) — одна из важных задач транскриптомного профилирования. Данные по дифференциальной экспрессии обычно представляются в матричном виде, где каждая строка соответствует гену (или транскрипту), а каждый столбец — образцу, в ячейках указывается уровень экспрессии гена в образце [2]. Основная исследовательская проблема — это обнаружение статистически значимых ДЭГ между различными группами образцов (например, здоровыми и больными). Одна из частых проблем, возникающих при статистической об-

Сокращения: ДЭГ — дифференциально экспрессирующиеся гены; ГКМП — гипертрофическая кардиомиопатия; ROC-AUC (ROC = receiver operating characteristic, AUC = area under the curve) — метрика качества классификации; p-val$_{MW}$ — p-value по критерию Манна–Уитни; FDR$_{BH}$ (FDR — False Discovery Rate) — поправка на множественные сравнения Бенджамини–Хохберга; FDR$_{wBH}$ — взвешенная поправка на множественные сравнения Бенджамини–Хохберга; Вес$_{ML}$ — вес гена, отображающий его значимость для классификационных моделей по результатам симуляций Монте–Карло; $\log_2$FC — логарифм отношения средних; ML (Machine Learning) — машинное обучение.

работке таких данных, связана с “проклятием размерности” [3].

“Проклятие размерности” — это феномен, при котором с увеличением количества измерений или переменных входных данных увеличивается объем пространства признаков, что может привести к увеличению шума и ошибочным выводам. Средняя размерность пространства признаков данных при транскриптомном профилировании превышает 10 000. Средний размер выборок меньше 100 точек. Таким образом, несмотря на богатство информации, получаемой с помощью высокопроизводительных методов исследования, интерпретация этих данных может быть сложной из-за большого количества генов и малого количества образцов.

К стандартным средствам решения обозначенной проблемы относятся различные инструменты корректировки значений *p*-value с учетом множественных сравнений, широко используемые внутри таких популярных пакетов, как EdgeR [4] или Limma [5].

В данной работе мы предлагаем новый подход, основанный на использовании методов машинного обучения (ML), для уменьшения размерности данных и выделения ключевых генов, имеющих наибольший шанс быть ассоциированными с исследуемым заболеванием, с последующим применением взвешенных процедур коррекции на множественные сравнения. Веса для корректировки значений *p*-value также получаются с помощью методов ML.

Суть подхода заключается в том, чтобы на данных транскриптомного профилирования разыграть методом Монте-Карло классификаторы с высокой обобщающей способностью. Далее из этих классификаторов извлекаются важные для их работы признаки, или ключевые гены и формируется редуцированное пространство признаков для последующего тестирования в нем гипотез об ассоциации стандартными методами. Полученное пространство признаков также будет взвешенным пространством, т.е. с заданной на нем весовой функцией или мерой. Вес будет задаваться как доля моделей, в которую был включен ген, умноженная на метрику качества ROC-AUC, усредненную по этим моделям. Этот вес будет использоваться при проведении взвешенных процедур коррекции на множественное тестирование гипотез, таких как взвешенные методы Бонферрони, Холма или Бенджамини–Хохберга.

Первоначально перечисленные методы взвешенной коррекции были разработаны для возможности учета априорной информации [6, 7]. В настоящее время бóльшая часть работ, посвященная развитию этих методов, сводится к постановке задачи максимизации мощности статистических тестов по вектору весов [8, 9]. В представленной работе мы предлагаем вернуться к классической постановке с заданием весовых коэффициентов, отражающих некоторую априорную информацию, которые мы получаем из данных (*data driven approach*), а именно из эффективности работы классификаторов. Другими словами, как описано выше, чем в бóльшее число хорошо работающих классификаторов включен тот или иной ген в Монте-Карло симуляциях, тем выше его вес.

Таким образом, в представленном исследовании вместо распространенного подхода (от фундаментальных наблюдений за изменениями транскриптома при различных состояниях к созданию классификатора для целей прикладной медицины), мы предлагаем идти в обратном направлении: от эффективно работающих классификаторов к пониманию патогенетических процессов, приводящих к изменениям в транскриптоме, которые и улавливаются этими классификаторами.

Для демонстрации работы предлагаемого подхода были выбраны открытые данные транскриптомного профилирования больных гипертрофической кардиомиопатией (ГКМП): GSE36961 и GSE1145.

МЕТОДЫ

Кратко, на первом этапе мы начинаем с загрузки и предварительной обработки набора данных GSE36961 по стандартному протоколу [5]. Для обучения классификаторов мы формируем матрицу данных размера $n \times m$, где n — число наблюдений, m — число признаков/генов; зависимая переменная представляет собой вектор из (0, 1), где 0 — отсутствие ГКМП, 1 — наличие ГКМП. Задача классификации ставится так, чтобы научиться по вектору признаков (уровней экспрессии генов) предсказывать “наличие ГКМП”.

Для поиска генов, вовлеченных в патогенез ГКМП, методом Монте-Карло мы разыгрывали L1-регуляризованные классификаторы на базе логистической регрессии. L1-регуляризация позволяет прореживать признаковое пространство, оставляя в классификационной модели только наиболее значимые признаки (гены). Используя это свойство, мы и будем осуществлять отбор признаков. Далее мы обучали 3000 моделей (про-

водили 3000 симуляций), извлекая обучающую выборку по схеме с возвращением. Извлекали только наблюдения (строки). Гены (признаки, столбцы) не извлекались. Каждое наблюдение (строка) извлекалось с возвращением равновероятно и независимо. Тестовая выборка формировалась из наблюдений, которые не попали в обучающую выборку. В итоге обучающая и тестовая выборка формировались в примерном соотношении 8 : 2. Таким образом, мы не полагаемся на одну модель, а симулируем множество различных экспериментов на различных выборках, полученных за счет извлечения исходной выборки.

Перед запуском алгоритма регуляризационный коэффициент подбирали так, чтобы качество модели по метрике ROC-AUC снижалось минимально. Подбор коэффициента и оценку качества осуществляли на размеченной обучающей выборке, используя кросс-валидацию. Таким образом, мы допускаем переобучение, но оставляем максимальное количество генов, исходя из идеи, что несостоятельные признаки будут реже включаться в модель, что напрямую отразиться на их весе.

Исходя из обученных моделей, составляли множество отобранных генов, которым присваивали вес по следующей формуле:

$$weight_{gene_j} = \frac{\sum\left(I_{gene_j \in model_i}\right)}{n} \cdot \frac{\sum\left(ROCAUC_i \cdot I_{gene_j \in model_i}\right)}{\sum\left(I_{gene_j \in model_i}\right)}, \quad (1)$$

где $I_{gene_j \in model_i}$ — индикатор включения j-го гена в i-ю модель, $ROCAUC_i$ — метрика ROC-AUC для i-модели, n— число итераций.

Таким образом, в качестве веса принята доля моделей, в которую был включен ген, умноженная на метрику качества ROC-AUC, усредненную по этим моделям. ROC-AUC модели включаются в расчет веса гена, чтобы различать гены, отобранные в одинаковое число моделей, но различающиеся качеством классификации. В последующем нас будут интересовать гены, которые чаще всего включаются в наилучшие классификаторы. В этом случае присваиваемый вес позволит релевантным образом упорядочить список генов для их последующей обработки. Гены, которые входят в состав менее, чем 5% моделей и имеют низкий вес, будут исключены из дальнейшего рассмотрения.

Валидацию результатов проводили на независимом наборе данных (GSE1145), который не использовали при обучении или тестировании. Для оценки ассоциации (тестирование гипотезы левого или правого сдвига) использовали непараметрический критерий Манна–Уитни [10], поправку на множественные сравнения Бенджамини–Хохберга [11], а также взвешенную поправку на множественные сравнения Бенджамини–Хохберга по схеме, описанной в работе [12].

Статистические тесты проводили с использованием модуля SciPy version: 1.7.3. Для обучения моделей, их тестирования и препроцессинга данных использовали модуль sklearn version: 0.24.2 [13].

Код алгоритма доступен по ссылке: https://github.com/GJOsmak/MolBiol2024.

РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ

Всего на чипе GSE36961 представлено 37846 транскриптов. После предобработки данных, удаления пропусков, мультимаперов и нулевых прочтений для анализа осталось 14830 транскриптов. Таким образом, изначальная размерность пространства составила 14830 при выборке объемом 145 наблюдений. На этих данных мы провели 3000 Монте-Карло симуляций, как это описано в разделе Методы.

Мы предполагаем, что если какой-то ген сильно ассоциирован с исследуемым заболеванием, то он будет входить в большинство моделей вне зависимости от способа разбиения выборки (номера итерации). При оценке сходимости алгоритма мы решили назвать “наиболее значимыми генами” те, которые включаются не менее, чем в половину моделей.

Как видно из рис. 2*а*, алгоритм сходится по числу наиболее значимых генов: после ~2000-ой итерации состав таких генов не меняется и сходится к шести генам (*MYH6*, *CDC42EP4*, *RASD1*, *PRKCD*, *FCN3*, *ZFP36*). Из рис. 2*б* видно, что после 2000 итераций прирост новых генов (зеленая линия), как и увеличение веса наиболее значимых генов (красная линия), выходят на стационарное состояние. При этом скорость увеличения веса наиболее значимых генов превосходит скорость прироста новых генов. Следовательно, можно предположить, что все ассоциированные с исследуемым заболеванием гены были отобраны на 2000 итерациях. Все отбираемые в

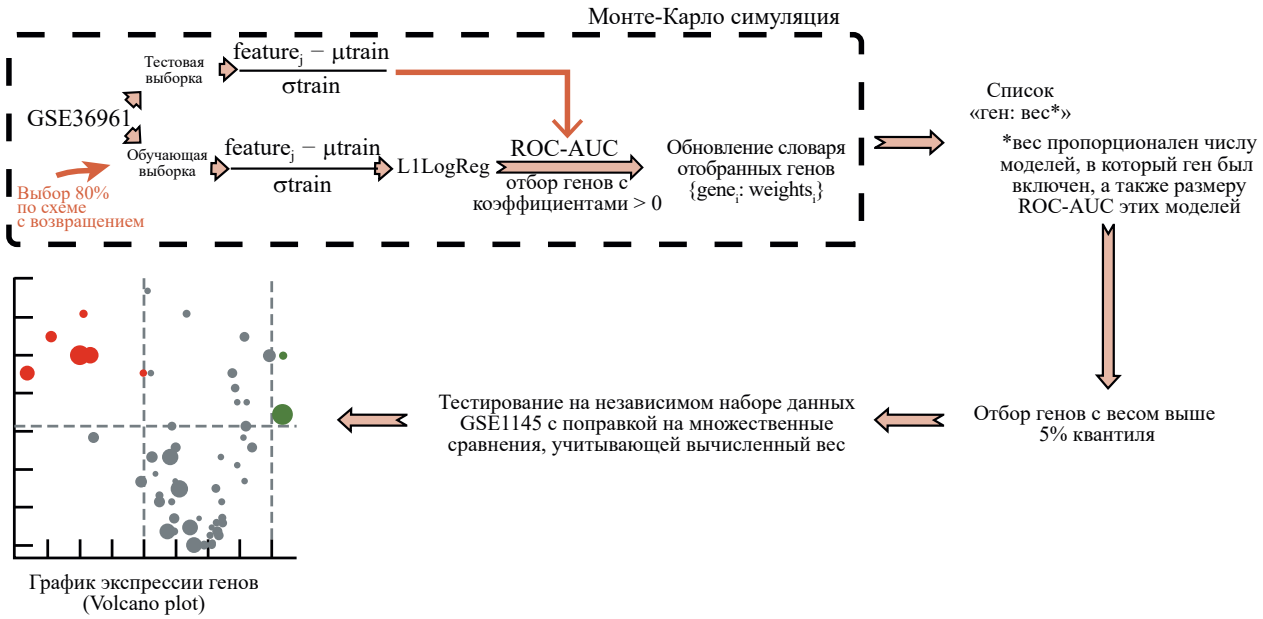


Рис. 1. Схема исследования.

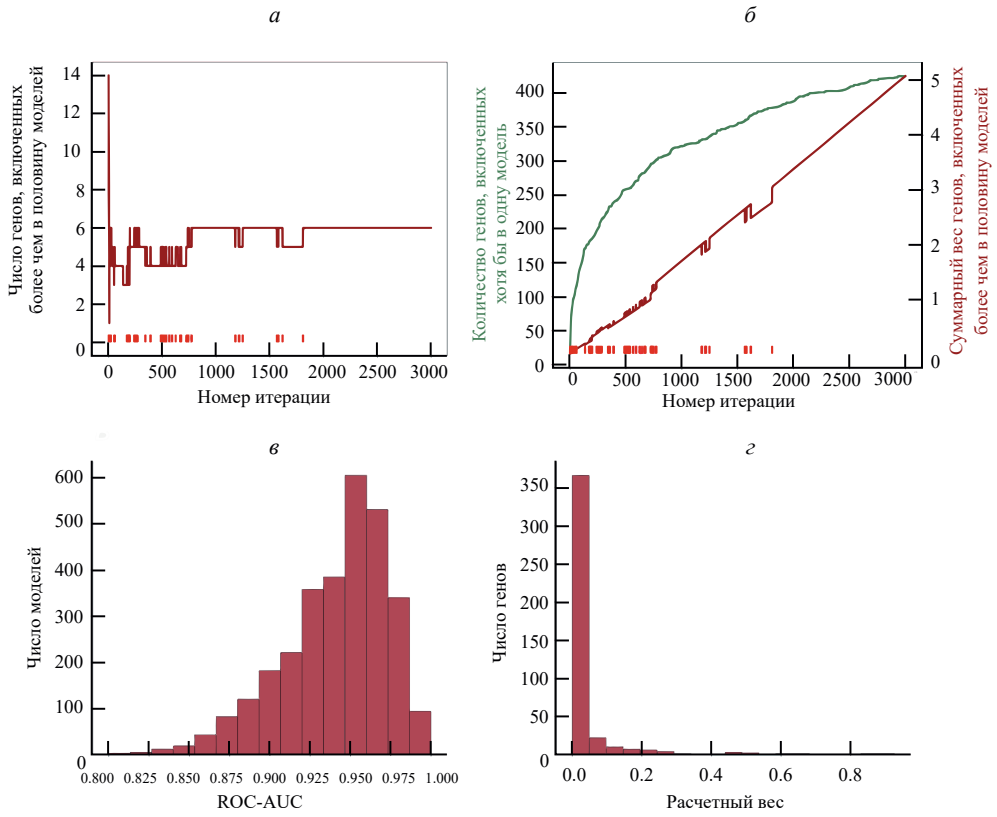


Рис. 2. Результаты проведения Монте–Карло симуляций по обучению классификаторов. *а* — Сходимость алгоритма по объему множества наиболее значимых генов; красные штрихи вдоль оси абсцисс показывают моменты изменения состава этого множества. *б* — Динамика роста в зависимости от итерации алгоритма числа отбираемых генов (зеленая линия); веса генов, включенных более чем в половину моделей (красная линия); итерация, на которой изменено множество наиболее значимых генов (красные вертикальные штрихи вдоль оси абсцисс). *в* — Гистограмма распределения меры ROC-AUC для ML-классификаторов в 3000 симуляциях Монте–Карло. *г* — Гистограмма распределения расчетного веса генов, включенных, по крайней мере, в одну модель.

последующем гены относятся к шуму и связаны скорее со способом разбиения выборки, чем с исследуемым заболеванием.

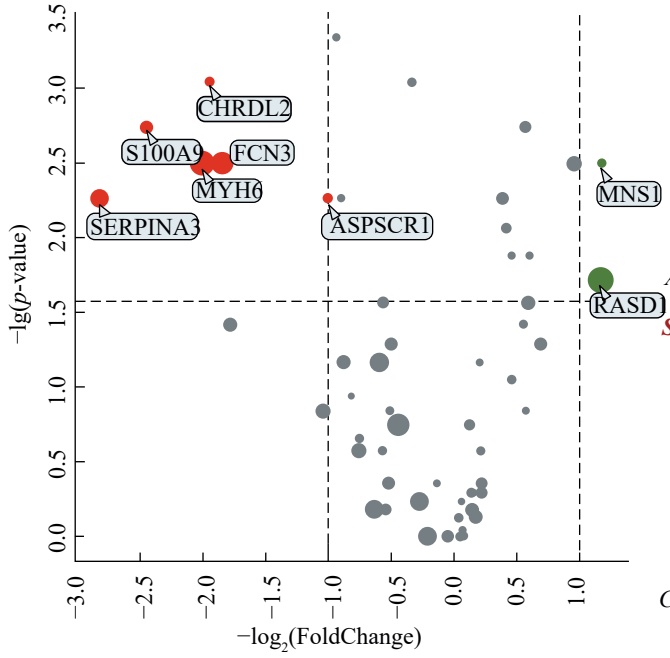
В результате, по меньшей мере в одну из 3000 моделей были включены 425 генов в различных комбинациях. Как видно из рис. 2*в*, большинство моделей обладают высоким показателем ROC-AUC (больше 0.9). При этом большинство генов (368 из 425) включаются меньше чем в 5% моделей (рис. 2*г*). Исходя из предположения, что ассоциированный с заболеванием ген будет входить в большинство моделей, мы делаем вывод, что полезность этих 368 генов для классификаторов связана скорее со способом разбиения выборки, чем с заболеванием. Для последующего анализа веса этих генов приравниваются к нулю. В результате пространство тестируемых гипотез сокращается до 57 генов, т.е. в 260 раз от объема изначального пространства (14830 генов).

Как видно из рис. 3, не все из отобранных выше шести “наиболее значимых генов” оказались статистически значимо ассоциированы с исследуемым заболеванием. Ассоциация не подтвердилась для генов *CDC42EP4*, *PRKCD*, *ZFP36*. С другой стороны, из рис. 3*а* видно, что наравне с генами *MYH6*, *FCN3* и *RASD1*, статистически значимо ассоциирован и сильно меняет свою экспрессию по $\log_2$FC ген *SERPINA3*, который не добрал 0.06 долей веса, чтобы войти в список “наиболее значимых генов”. Из рис. 3*б* видно, что не все гены, прошедшие поправку на множественные сравнения (FDR_{BH}), прошли взвешенную поправку (FDR_{wBH}). К таким генам относятся *INTU, HEG1, SYF2, NKD2, ASPSCR1.*

ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ

В рамках данного исследования мы разработали и успешно применили алгоритм на основе метода Монте-Карло для разыгрывания устойчивых классификаторов и прореживания с их помощью пространства признаков (генов). В результате этого анализируемое пространство было уменьшено в ~260 раз с 14830 до 57 генов, которые при последующем тестировании гипотез об ассоциации сократились до 12 генов: *MNS1, FCN3, CHRDL2, MYH6, CAPN1, CD97, S100A9, PROS1, CHN1, SERPINA3, AP3M2, RASD1*, из которых по совокупности признаков (расчетный вес, $\log_2$FC, скорректированный

а



б

Ген	p-val$_{MW}$	FDR$_{BH}$	FDR$_{wBH}$	Вес$_{ML}$	$\log_2$FC
CD97	**0.000**	**0.017**	**0.017**	**0.064**	**−0.94**
CHRDL2	**0.001**	**0.017**	**0.017**	**0.118**	**−1.95**
CAPN1	**0.001**	**0.017**	**0.017**	**0.098**	**−0.33**
S100A9	**0.002**	**0.02**	**0.017**	**0.234**	**−2.45**
CHN1	**0.002**	**0.02**	**0.017**	**0.159**	**0.57**
MNS1	**0.003**	**0.02**	**0.033**	**0.097**	**1.18**
FCN3	**0.003**	**0.02**	**0.017**	**0.621**	**−1.84**
MYH6	**0.003**	**0.02**	**0.017**	**0.832**	**−2.01**
PROS1	**0.003**	**0.02**	**0.017**	**0.322**	**0.96**
ASPSCR1	0.005	0.024	0.06	0.086	−1.01
NKR2	0.005	0.024	0.064	0.069	−0.9
SERPINA3	**0.005**	**0.024**	**0.017**	**0.443**	**−2.82**
AP3M2	**0.005**	**0.024**	**0.033**	**0.183**	**0.39**
SYF2	0.009	0.035	0.06	0.126	0.42
INTU	0.013	0.047	0.131	0.058	0.6
HEG1	0.013	0.047	0.115	0.071	0.46
RASD1	**0.019**	**0.064**	**0.025**	**0.926**	**1.17**
JAK2	0.027	0.082	0.088	0.236	0.59
S1PR3	0.027	0.082	0.115	0.144	−1.56
MGST1	0.038	0.103	0.097	0.263	−1.78
ZNF415	0.038	0.103	0.222	0.077	−0.55
PRKCD	0.068	0.144	0.092	0.526	−0.59
CDC42EP4	0.180	0.304	0.134	0.708	−0.45
ZFP36	0.661	0.750	0.460	0.518	−0.63

Рис. 3. Тестирование гипотез об ассоциации отобранных генов на независимом наборе данных GSE1145. *а* — График сравнения экспрессии генов (Volcano plot), размер точек обозначает их Вес$_{ML}$. *б* — Сводная таблица статистик; показаны только значимые (по *p*-value) результаты. *p*-val$_{MW}$ — *p*-value по критерию Манна–Уитни; FDR$_{BH}$ — поправка на множественные сравнения Бенджамини–Хохберга; FDR$_{wBH}$ — взвешенная поправка на множественные сравнения Бенджамини–Хохберга; Вес$_{ML}$ — вес гена, отображающий его значимость для классификационных моделей по результатам Монте–Карло симуляций; $\log_2$FC — логарифм отношения средних.

p-value) наибольшего внимания заслуживают *MYH6*, *FCN3*, *RASD1* и *SERPINA3*.

Большая часть моделей при обучении обладала высокими показателями метрики ROC-AUC (мода = 0.96, см. рис. 2*в*). С другой стороны, большинство генов включались меньше чем в 5% моделей (см. рис. 2*г*). Этот результат хорошо согласуется с последствиями обучения моделей в пространстве высокой размерности, где легко подобрать такой набор признаков, в пространстве которых конкретно взятая выборка будет хорошо разделима, однако это будет артефактом, а не ценным результатом [3].

Ген *MYH6* кодирует альфа-изоформу тяжелой цепи сердечного миозина (α-MHC), которая экспрессируется во всем миокарде на ранних стадиях развития сердца. По мере развития эмбриона человека экспрессия гена *MYH6* в желудочках снижается и заменяется экспрессией *MYH7* [14]. В ряде работ показана ассоциация этого гена с ГКМП [15, 16].

Продукт гена *FCN3* — мощный активатор пути комплемента на основе лектина [17], ассоциированный, согласно [18, 19], с сердечной недостаточностью и ишемической кардиомиопатией [20].

Мономерный белок RASD1 экспрессируется в сердечной ткани на низком уровне [21]. Нокдаун гена *RASD1* в кардиомиоцитах предсердий, приводит к существенному увеличению экспрессии предсердного натрийуретического фактора [22, 23], однако каких-либо связей RASD1 с кардиомиопатиями на данный момент не выявлено.

SERPINA3, также называемый α-1-антихимотрипсином (AACT, ACT), является одним из ингибиторов сериновых протеаз, в частности катепсина G [24]. Как белок острой фазы, секретируемый в плазму клетками печени, SERPINA3 играет важную роль в противовоспалительной реакции и противовирусном ответе. Повышенные уровни SERPINA3 наблюдаются при сердечной недостаточности и неврологических заболеваниях [25].

Таким образом, часть обнаруженных с помощью предложенного алгоритма генов непосредственно связана с исследуемым заболеванием, другая часть — косвенно, т.е. полученные результаты не противоречат опубликованным данным. Стоит также отметить, что такие же наборы данных, GSE36961 и GSE1145, анализируют в работе [26], используя “стандартные” подходы, и приходят к похожему набору генов: *RASD1, CDC42EP4, MYH6* и *FCN3*. Таким образом, предлагаемый нами подход хорошо соответствует результатам стандартных подходов, а его преимущество состоит в возможности полной алгоритмизации и минимальном количестве произвольных решений. В дополнение, по результатам нашего анализа добавляется еще один параметр оценки “значимости” генов — вес. Варианты его использования показаны на рис. 3.

ЗАКЛЮЧЕНИЕ

В нашей работе предложен новый алгоритм анализа данных транскриптомного профилирования. Результаты работы алгоритма хорошо согласуются с опубликованными данными и открывают новые возможности анализа посредством генерации взвешенного пространства признаков (генов), в противовес “стандартной” ситуации, когда все признаки (гены) рассматриваются как “равные”.

Работа поддержана грантом РНФ № 23-75-01050. Работа выполнена без привлечения людей и животных в качестве объектов исследования.

Авторы заявляют об отсутствии конфликта интересов.

СПИСОК ЛИТЕРАТУРЫ

1. Akond Z., Alam M., Mollah Md.N.H. (2018) Biomarker identification from RNA-seq data using a robust statistical approach. *Bioinformation*. **14**(4), 153–163.
2. Tang M., Sun J., Shimizu K., Kadota K. (2015) Evaluation of methods for differential expression analysis on multi-group RNA-seq count data. *BMC Bioinformatics*. **16**(1), 360.
3. Barbiero P., Squillero G., Tonda A. (2020) Modeling generalization in machine learning: a methodological and computational study. *arXiv*. **2006**.15680.
4. Robinson M.D., McCarthy D.J., Smyth G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. **26**(1), 139–140.
5. Smyth G.K. (2005) Limma: linear models for microarray data. In: *Bioinformatics and computational biology solutions using R and Bioconductor*. New York: Springer.
6. Benjamini Y., Hochberg Y. (1997) Multiple hypotheses testing with weights. *Scandinavian J. Statistics*. **24**(3), 407–418.
7. Holm S. (1979) A simple sequentially rejective multiple test procedure. *Scandinavian J. Statistics*. **6**(2), 65–70.
8. Gui J., Tosteson T.D., Borsuk M. (2012) Weighted multiple testing procedures for genomic studies. *BioData Mining*. **5**(1), 4.

9. Basu P., Cai T. T., Das K., Sun W (2018) Weighted false discovery rate control in large-scale multiple testing. *J. Am. Stat. Assoc.* **113**(523), 1172–1183.
10. Mann H.B., Whitney D.R. (1947) On a test of whether one of two random variables is stochastically larger than the other. *Ann. Mathemat. Statistics.* **18**(1), 50–60.
11. Benjamini Y., Hochberg Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Royal Statist. Soc.: Series B (Methodological)*. **57**(1), 289–300.
12. Genovese C.R., Roeder K., Wasserman L. (2006) False discovery control with *p*-value weighting. *Biometrika*. **93**(3), 509–524.
13. Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Duchesnay E. (2011) Scikit-learn: machine learning in python. *J. Machine Learning Res.* **12**(Oct), 2825–2830.
14. Anfinson M., Fitts R.H., Lough J.W., James J.M., Simpson P.M., Handler S.S., Mitchell M.E., Tomita-Mitchell A. (2022) Significance of α-myosin heavy chain (MYH6) variants in hypoplastic left heart syndrome and related cardiovascular diseases. *J. Cardiovascular Dev. Dis.* **9**(5), 144.
15. Ntelios D., Meditskou S., Efthimiadis G., Pitsis A., Zegkos T., Parcharidou D., Theotokis P., Alexouda S., Karvounis H., Tzimagiorgis G. (2022) α-Myosin heavy chain (MYH6) in hypertrophic cardiomyopathy: prominent expression in areas with vacuolar degeneration of myocardial cells. *Pathol. Int.* **72**(5), 308–310.
16. Suzuki T., Saito K., Yoshikawa T., Hirono K., Hata Y., Nishida N., Yasuda K., Nagashima M. (2022) A double heterozygous variant in *MYH6* and *MYH7* associated with hypertrophic cardiomyopathy in a Japanese family. *J. Cardiol. Cases.* **25**(4), 213–217.
17. Michalski M., Świerzko A.S., Pągowska-Klimek I., Niemir Z.I., Mazerant K., Domżalska-Popadiuk I., Moll M., Cedzyński M. (2015) Primary ficolin-3 deficiency — is it associated with increased susceptibility to infections? *Immunobiology.* **220**(6), 711–713.
18. Prohászka Z., Munthe-Fog L., Ueland T., Gombos T., Yndestad A., Förhécz Z., Skjoedt MO, Pozsonyi Z., Gustavsen A., Jánoskuti L., Karádi I., Gullestad L., Dahl C.P., Askevold E.T., Füst G., Aukrust P., Mollnes T.E., Garred P. (2013) Association of ficolin-3 with severity and outcome of chronic heart failure. *PLoS One.* **8**(4), e60976.
19. Li D., Lin H., Li L. (2020) Multiple feature selection strategies identified novel cardiac gene expression signature for heart failure. *Front. Physiol.* **11**, 604241.
20. Song H., Chen S., Zhang T., Huang X., Zhang Q., Li C., Chen C., Chen S., Liu D., Wang J., Tu Y., Wu Y., Liu Y. (2022) Integrated strategies of diverse feature selection methods identify aging-based reliable gene signatures for ischemic cardiomyopathy. *Front. Mol. Biosci.* **9**, 805235.
21. Wie J., Kim B.J., Myeong J., Ha K., Jeong S.J., Yang D., Kim E., Jeon J.H., So I. (2015) The roles of Rasd1 small G proteins and leptin in the activation of TRPC4 transient receptor potential channels. *Channels.* **9**(4), 186–195.
22. Kemppainen R.J., Behrend E.N. (1998) Dexamethasone rapidly induces a novel *Ras* superfamily member-related gene in AtT-20 cells. *J. Biol. Chem.* **273**(6), 3129–3131.
23. McGrath M.F., Ogawa T., De Bold A.J. (2012) Ras dexamethasone-induced protein 1 is a modulator of hormone secretion in the volume overloaded heart. *Am. J. Physiol. Heart Circ. Physiol.* **302**(9), H1826–H1837.
24. Baker C., Belbin O., Kalsheker N., Morgan K. (2007) SERPINA3 (aka alpha-1-antichymotrypsin). *Front. Biosci.* **12**(8–12), 2821–2835.
25. de Mezer M., Rogaliński J., Przewoźny S., Chojnicki M., Niepolski L., Sobieska M., Przystańska A. (2023) SERPINA3: stimulator or inhibitor of pathological changes. *Biomedicines.* **11**(1), 156.
26. You H., Dong M. (2023) Prediction of diagnostic gene biomarkers for hypertrophic cardiomyopathy by integrated machine learning. *J. Int. Med. Res.* **51**(11), 03000605231213781.

Transcriptomics and the “Curse of Dimensionality”: Monte Carlo Simulations of ML-Models as a Tool for Analyzing Multidimensional Data in Tasks of Searching Markers of Biological Processes

© 2025 г. G. J. Osmak*, M. V. Pisklova

Chazov National Medical Research Center for Cardiology, Moscow, 121552 Russia
Pirogov Russian National Research Medical University, Moscow, 117997 Russia
**e-mail: german.osmak@gmail.com*

High-throughput transcriptomic research methods provide the assessment of a vast number of factors, valuable for researchers. At the same time the “curse of dimensionality” issues arise, which lead to increasing requirements on data processing and analysis methods. In this study, we propose a new algorithm that combines Monte Carlo methods and machine learning. This algorithm will enable feature space reduction by highlighting genes most likely associated with the investigated diseases. Our approach allows not only to generate a set of “interesting” genes but also to assign weight to each gene, indicating its “importance”. This measure can be used in subsequent statistical analysis, visualization, and interpretation of results. Algorithm performance was demonstrated on open transcriptomic data of patients with HCM (GSE36961 and GSE1145). The analysis revealed genes *MYH6, FCN3, RASD1*, and *SERPINA3*, which is in good agreement with the available literature.

Keywords: transcriptomics, machine learning, Monte Carlo, hypertrophic cardiomyopathy, biomarkers