НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
май–июнь 2025 Том 25 № 3 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
May–June 2025 Vol. 25 No 3 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

МАТЕМАТИЧЕСКОЕ И КОМПЬЮТЕРНОЕ МОДЕЛИРОВАНИЕ
MODELING AND SIMULATION

doi: 10.17586/2226-1494-2025-25-3-487-497
УДК 004.961

Метод определения активного модуля в биологических графах с многокомпонентными весами вершин

Дмитрий Андреевич Усольцев[1]✉, Иван Игоревич Молотков[2], Никита Николаевич Артемов[3], Алексей Александрович Сергушичев[4], Анатолий Абрамович Шалыто[5]

[1,2,3] Институт геномной медицины, Детская больница Нейшенвайд, Колумбус, 43205, США
[1,5] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация
[2,3] Медицинский колледж Университета штата Огайо, Колумбус, 43210, США
[4] Университет Вашингтона в Сент-Луисе, Сент-Луис, 63110, США

[1] dusoltsev.27@gmail.com✉, https://orcid.org/0000-0001-8072-310X
[2] ivan.molotkov@nationwidechildrens.org, https://orcid.org/0009-0008-3566-0160
[3] mykyta.artomov@nationwidechildrens.org, https://orcid.org/0000-0001-5282-8764
[4] asergushichev@wustl.edu, https://orcid.org/0000-0003-1159-7220
[5] anatoly.shalyto@gmail.com, https://orcid.org/0000-0002-2723-2077

Аннотация

Введение. Активный модуль в биологических графах представляет собой связанный подграф, вершины которого объединены общей биологической функцией. Для определения активного модуля необходимо сначала построить взвешенный биологический граф. Вес каждой вершины вычисляется на основе биологических экспериментов, исследующих искомую биологическую функцию. Однако результаты одного эксперимента могут не полностью описывать искомый активный модуль, а лишь его часть, внося, например, неопределенность в веса вершин. В работе показано, что использование метода Фишера для объединения данных нескольких экспериментов, а затем применение метода Монте-Карло по схеме марковских цепей (МКМЦ) и машинного обучения к результатам метода Фишера, позволяет более эффективно выделять активные модули в биологических графах. **Метод.** В работе используются граф белок-белковых взаимодействий — InWebIM, граф по реконструкции мозга человека из проекта BigBrain и генный граф для вида живых организмов *Caenorhabditis elegans*. Для объединения результатов нескольких экспериментов в одном графе в единый результат применяется метод Фишера. После этого поиск активных модулей выполняется с использованием метода МКМЦ и машинного обучения. Для валидации предлагаемого метода на реальных данных применяются результаты полногеномного ассоциативного исследования по шизофрении и курению, а также матрица экспрессии генов пациентов с кожной меланомой из проекта The Cancer Genome Atlas. **Основные результаты.** Применение метода Фишера позволяет учитывать результаты нескольких биологических экспериментов одновременно. Последующее использование метода МКМЦ и машинного обучения повышает точность определения активных модулей по сравнению с ранжированием вершин графа только на основе метода Фишера. **Обсуждение.** Учет результатов нескольких биологических экспериментов при определении активных модулей играет ключевую роль в повышении точности нахождения вершин активного модуля. Это способствует лучшему пониманию биологических механизмов заболеваний, что может иметь важное значение для разработки новых методов диагностики и терапии.

Ключевые слова

графы, метод Монте-Карло по схеме марковских цепей, метод Фишера, биологические графы, активный модуль

Ссылка для цитирования: Усольцев Д.А., Молотков И.И., Артемов Н.Н., Сергушичев А.А., Шалыто А.А. Метод определения активного модуля в биологических графах с многокомпонентными весами вершин // Научно-технический вестник информационных технологий, механики и оптики. 2025. Т. 25, № 3. С. 487–497. doi: 10.17586/2226-1494-2025-25-3-487-497

Method for identifying the active module in biological graphs with multi-component vertex weights

Dmitrii A. Usoltsev[1]✉, Ivan I. Molotkov[2], Mykyta N. Artomov[3], Alexey A. Sergushichev[4], Anatoly A. Shalyto[5]

[1,2,3] Institute for Genomic Medicine, Nationwide Children's Hospital, Columbus, 43205, USA
[1,5] ITMO University, Saint Petersburg, 197101, Russian Federation
[2,3] The Ohio State University College of Medicine, Columbus, 43210, USA
[4] Washington University School of Medicine in St. Louis, St. Louis, 63110, USA

[1] dusoltsev.27@gmail.com✉, https://orcid.org/0000-0001-8072-310X
[2] ivan.molotkov@nationwidechildrens.org, https://orcid.org/0009-0008-3566-0160
[3] mykyta.artomov@nationwidechildrens.org, https://orcid.org/0000-0001-5282-8764
[4] asergushichev@wustl.edu, https://orcid.org/0000-0003-1159-7220
[5] anatoly.shalyto@gmail.com, https://orcid.org/0000-0002-2723-2077

Abstract

An active module in biological graphs is a connected subgraph whose vertices share a common biological function. To identify an active module, one must first construct a weighted biological graph. The weight of each vertex is calculated based on biological experiments investigating the target biological function. However, the results of a single experiment may not fully describe the desired active module, covering only part of it and potentially introducing uncertainty into the vertex weights. This work demonstrates that employing Fisher's method to integrate data from multiple experiments followed by applying a Markov chain Monte Carlo (MCMC) and machine learning–based approach to the results of Fisher's method, enables more effective identification of active modules in biological graphs. The study utilizes the InWebIM protein–protein interaction graph, a human brain reconstruction graph from the BigBrain project, and a gene graph for the organism *Caenorhabditis elegans*. To combine the results of several experiments into a single outcome within one graph, Fisher's method is applied. Afterwards, the search for active modules is conducted using an MCMC and machine learning-based method. To validate the proposed method on real data, results from Genome-Wide Association Studies on schizophrenia and smoking are used, along with the gene expression matrix of patients with skin melanoma from the TCGA project. Applying Fisher's method makes it possible to consider the results of multiple biological experiments simultaneously. Subsequent use of the MCMC and machine learning–based method improves the accuracy of identifying active modules compared to ranking graph vertices solely by Fisher's method. Considering the results of multiple biological experiments when determining active modules plays a crucial role in increasing the accuracy of identifying the vertices of the active module. This, in turn, promotes a deeper understanding of the biological mechanisms of diseases, which can be of great significance for the development of new diagnostic and therapeutic methods.

Keywords

graphs, MCMC, Fisher's method, biological graphs, active module

For citation: Usoltsev D.A., Molotkov I.I., Artomov M.N., Sergushichev A.A., Shalyto A.A. Method for identifying the active module in biological graphs with multi-component vertex weights. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2025, vol. 25, no. 3, pp. 487–497 (in Russian). doi: 10.17586/2226-1494-2025-25-3-487-497

Введение

Графы часто используются при анализе сложных биологических взаимодействий, таких как белок-белковые взаимодействия, ко-экспрессия генов или регуляция метаболических путей [1–3]. Внутри таких графов существуют связанные подграфы, вершины которых выполняют общую биологическую функцию. Эти подграфы называют активными модулями [4]. Модули, например, в графе белок-белковых взаимодействий позволяют понять причины возникновения заболеваний и определить возможные терапевтические мишени [5, 6]. Таким образом, определение активных модулей является одной из важных задач при анализе биологических графов.

В работах [7, 8] представлены подходы для определения активных модулей на основе метода Монте-Карло по схеме марковских цепей (МКМЦ) и метода на основе машинного обучения — градиентного бустинга. Однако точность определения активного модуля как с помощью только метода МКМЦ [7], так и с помощью подхода, предложенного в [8], зависит от точности определения весов вершин графа.

Для определения весов в биологическом графе проводятся эксперименты по дифференциальной экспрессии генов или мутационной нагрузке между больными и здоровыми группами людей. Результаты экспериментов анализируются с помощью статистического теста (например, t-тест, линейная регрессия) [9]. В итоге формируется список генов с соответствующими p-значениями, отражающими значимость ассоциации с исследуемым признаком. Распределение p-значений хорошо аппроксимируется бета-равномерным распределением BUM (α, λ) с плотностью, вычисляемой на основе соотношения:

$$f(x) = \lambda + (1 - \lambda)\alpha x^{\alpha-1},\ 0 \le x, \lambda, \alpha \le 1, \qquad (1)$$

где λ — вес равномерной компоненты; α — параметр формы бета-компоненты [10].

В работах [4, 9] предложено, исходя из p-значений, вычислять априорный вес вершины для заданного уровня ложноположительных результатов (False Discovery Rate, FDR) на основе соотношения:

$$w(p) = \left(\frac{p}{p\mathrm{FDR}}\right)^{\alpha-1},\ 0 \le p, \alpha \le 1, \qquad (2)$$

где p — p-значение; pFDR — вероятность того, что p-значение принадлежит равномерному распределению. Эта вероятность вычисляется по формуле [4, 9]:

$$p\text{FDR} = \left(\frac{(\lambda + (1-\lambda)\alpha - \text{FDR} \times \lambda}{\text{FDR}(1-\lambda)} \right)^{\left(\frac{1}{\alpha - 1}\right)}, \quad 0 \le \lambda, \alpha \le 1. \tag{3}$$

Традиционно веса вершин рассчитываются на основе результатов одного статистического теста по формулам (2) и (3). Однако в исследованиях часто доступны результаты множества статистических тестов, отражающих разные аспекты данных: дифференциальная экспрессия, эпигенетические модификации, ассоциации с фенотипами. Результаты каждого теста распределены на основе формулы (1). Учет нескольких статистических тестов для определения априорного веса вершин графа может позволить повысить точность анализа. Такой вес является многокомпонентным (рис. 1).

Для совмещения результатов нескольких статистических тестов используем метод Фишера, который включает вычисление статистики хи-квадрат:

$$X^2 = -2 \sum_{i=1}^{n} \ln(p_i), \; 0 < p_i \le 1, \tag{4}$$

где p_i — p-значение i-го эксперимента; n — число экспериментов (статистических тестов) (рис. 1).

Для вычисления вероятности получения статистики X^2 из распределения хи-квадрат (χ^2) со степенью свободы равной удвоенному числу экспериментов [11] используется соотношение:

$$p_{comb} = P(\chi^2 \ge X^2, df = 2 \times n), \tag{5}$$

где p_{comb} — результирующие p-значение; P — вероятность; df — степень свободы.

Активные модули часто достоверно не известны. Кроме того, один эксперимент позволяет определить лишь часть активного модуля. В этих условиях моделирование активных модулей с заданными параметрами в реальных биологических графах является оптимальным вариантом.

В настоящей работе выполнено моделирование активных модулей с заданными параметрами на подграфах с числом вершин $N = 1000$ в белок-белковом графе InWebIM [12]. Предложен метод определения активного модуля для случая, когда известны результаты n экспериментов, каждый из которых покрывает лишь часть активного модуля. Предлагаемый метод основан на последовательном использовании метода Фишера, метода МКМЦ и машинного обучения (используется градиентный бустинг[1]). Эффективность предлагаемого метода продемонстрирована сравнением его с простым ранжированием результатов, полученных только методом Фишера для разного числа экспериментов. Показана применимость предлагаемого метода в графе из проекта по реконструкции мозга BigBrain [13, 14] и генном графе (для вида живых организмов *Caenorhabditis elegans* (HS-LC) [13, 15]. Оценено качество метода Фишера и предлагаемого метода на реальных данных полногеномного ассоциативного исследования (Genome-Wide Association Study, GWAS) по шизофрении, курению и экспрессии генов пациентов с кожной меланомой.

[1] Chen T., He T., Benesty M., Khotilovich V., Tang Y., Cho H., Chen K., Mitchell R., Cano I., Zhou T., Li M., Xie J., Lin M., Geng Y., Li Y. xgboost: Extreme Gradient Boosting. R package version 1.5.0.2. [Электронный ресурс]. Режим доступа: https://CRAN.R-project.org/package=xgboost, свободный. Яз. англ. (дата обращения: 17.08.2024).

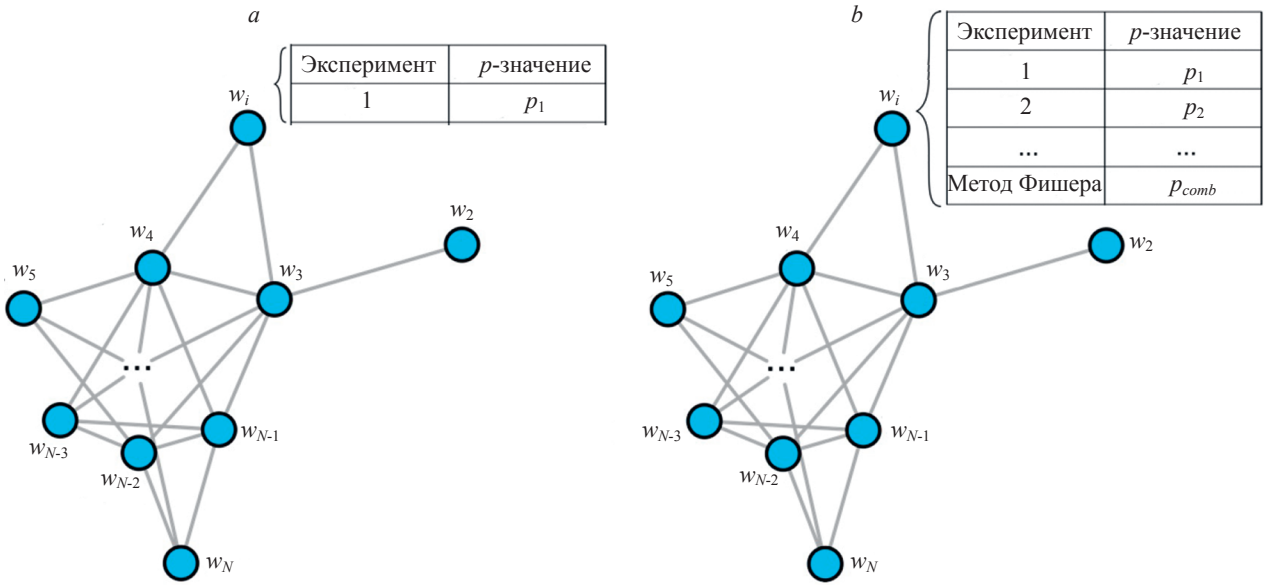

Рис. 1. Сравнение однокомпонентных (*a*) и многокомпонентных (*b*) весов вершин.
w_i — вес однокомпонентной i-ой вершины; w_j — вес многокомпонентной j-ой вершины; $i, j \in \{1, N\}$

Fig. 1. Comparison of single-component (*a*) and multi-component (*b*) node weights.
w_i — weight of the i-th single-component node; w_j — weight of the j-th multi-component node; $i, j \in \{1, N\}$

Предлагаемый метод

Этапы постановки математической задачи

1. **Определение активного модуля.** В отсутствии достоверно известных активных модулей для большинства биологических состояний проводится симуляция активных модулей с G вершинами в белок-белковом графе InWebIM посредством поиска в ширину из одной случайной вершины. Для повышения скорости вычислений этот граф сэмплируется (разбивается) на связанные подграфы — упрощенные белок-белковые графы. Число вершин в каждом подграфе $N = 1000$.

2. **Определение весов активного модуля.** Для вершин активного модуля генерируются p-значения из бета-распределения, для остальных вершин — из равномерного распределения. Генерация повторяется n раз. Для того чтобы промоделировать направленность биологических экспериментов на исследование только части активного модуля, на каждом этапе генерации p-значений 90 % вершин маскируются случайными значениями из равномерного распределения $X \sim U[0, 1]$. Таким образом, каждой вершине приписывается n независимых p-значений, что соответствует n экспериментам. С помощью метода Фишера для каждой вершины вычисляется одно p-значение. Затем вычисляется априорное значение принадлежности вершины к активному модулю.

Поиск активного модуля и оценка качества поиска

Поиск активного модуля выполняется с помощью метода, предложенного в [8], с учетом 20 % известных белков активного модуля. Качество определения активного модуля оценивается на основе метрик «площадь под кривой ошибок» (Receiver Operating Characteristic Area Under the Curve, ROC AUC; эта величина варьируется в диапазоне от нуля до единицы) и «чувствительность» (Recall@100 — доля правильно определенных белков активного модуля в первых 100 вершинах, ранжированных по уменьшению предсказанной с помощью модели вероятности вхождения в активный модуль; эта величина варьируется в диапазоне от нуля до единицы).

Этапы решения задачи

1. Симуляция 100 различных графов с известными активными модулями

1.1. Для получения графов с топологией, встречающейся в реальных данных, выбираются случайные подграфы из графа InWebIM с числом вершин равным 1000.

1.2. В каждом подграфе InWebIM из п. 1.1 равновероятно среди всех возможных связных подграфов заданного размера выбирается активный модуль с использованием метода МКМЦ из mcmcRanking (v0.1.0)[1] при условии, что все вершины имеют одинаковый вес. Модули выбираются после 1000 итераций МКМЦ, чтобы гарантировать их независимость от подграфов, использованных для инициализации МКМЦ.

2. Определение весов активного модуля

2.1. Для каждой вершины симулируются p-значения эксперимента, проверяющего принадлежность вершины активному модулю. Для вершин вне активного модуля p-значения выбираются из равномерного распределения, для вершин из активного модуля — из бета-распределения. Для 90 % вершин активного модуля p-значения маскируются случайными значениями из равномерного распределения на интервале (0, 1].

2.2. П. 2.1 повторяется n раз, таким образом, чтобы каждой вершине соответствовало n p-значений.

2.3. Для каждого p-значения применяется метод Фишера: считается X^2 по формуле (4) и результирующее p-значение по формуле (5).

2.4. Полученное p-значение аппроксимацией бета-равномерным распределением преобразуется в априорный вес по формулам (2) и (3).

3. Поиск активного модуля

3.1. Вероятность того, что подграф является активным модулем, рассчитывается как произведение вероятностей того, что каждая вершина принадлежит этому модулю. Для получившегося вероятностного пространства на множестве связных подграфов формируется выборка размера 100 с помощью метода МКМЦ из mcmcRanking (v0.1.0), используя 10 000 симуляций МКМЦ. После этого для каждой вершины определяется эмпирическая вероятность вхождения в активный модуль — доля подграфов из выборки, которые включают в себя эту вершину.

3.2. Применяется метод, изложенный в [8]. 20 % вершин активного модуля выбираются с помощью генератора случайных чисел и используются как известные вершины. Вычисляются расстояния от каждой вершины до трех ближайших белков активного модуля. Используя полученные расстояния и вероятность, определенную в п. 3.1, обучается модель градиентного бустинга с параметрами, описанными в [8]. Так как общее число активных модулей равняется 100, то 50 используются для тренировки модели, а оставшиеся 50 — для тестирования. Известные белки в активных модулях исключаются. При тренировке модели белкам, включенным в активный модуль, присваивается единица, а остальным белкам — ноль.

Экспериментальная оценка предлагаемого метода

Симуляции были проведены для $G = 100$ и $n \in \{1, 3, 10, 50\}$. Эффективность метода Фишера и метода, изложенного в [8], оценивалась с помощью метрик ROC AUC и Recall@100. Метрики определялись для каждого подграфа InWebIM, после этого вычислялось среднее значение каждой метрики из выборки в 50 подграфов. При $n = 1$ результаты, получаемые предлагаемым методом и методом, изложенным в [8] совпали, при условии маскировки p-значений для 90 % вершин активного модуля случайными значениями из равномерного распределения на интервале (0, 1].

[1] GitHub — ctlab/mcmcRanking: Tool To Solve The Active Module Problem [Электронный ресурс]. Режим доступа: https://github.com/ctlab/mcmcRanking, свободный. Яз. англ. (дата обращения: 15.12.2024).

Таблица. Сравнение предлагаемого метода и метода Фишера в зависимости от параметров α и *n*
Table. Comparison of the proposed method and Fisher's method depending vs. the parameters α и *n*

Параметры		Предлагаемый метод		Метод Фишера	
α	*n*	ROC AUC (±std)	Recall@100 (±std)	ROC AUC (±std)	Recall@100 (±std)
0,2	1	**0,69 ± 0,11**	**0,38 ± 0,19**	**0,54 ± 0,03**	**0,15 ± 0,03**
0,2	3	0,68 ± 0,09	0,36 ± 0,17	0,58 ± 0,03	0,21 ± 0,04
0,2	10	0,74 ± 0,07	0,46 ± 0,13	0,70 ± 0,03	0,36 ± 0,04
0,2	50	**0,87 ± 0,05**	**0,70 ± 0,11**	**0,90 ± 0,02**	**0,68 ± 0,06**
0,8	1	**0,61 ± 0,12**	**0,26 ± 0,19**	**0,51 ± 0,04**	**0,12 ± 0,03**
0,8	3	0,65 ± 0,11	0,30 ± 0,21	0,51 ± 0,04	0,12 ± 0,05
0,8	10	0,62 ± 0,10	0,27 ± 0,16	0,52 ± 0,03	0,11 ± 0,03
0,8	50	**0,67 ± 0,14**	**0,37 ± 0,24**	**0,54 ± 0,03**	**0,14 ± 0,04**

Примечание. (±std) — стандартное отклонение. Жирным шрифтом выделены значения для $n = 1$ и $n = 50$.

В таблице представлены результаты качества определения вершин активного модуля предлагаемым методом и методом Фишера для $\alpha = 0{,}2$ и $\alpha = 0{,}8$ и число экспериментов *n*.

При увеличении *n* в случае $\alpha = 0{,}2$, когда бета-равномерное распределение содержит более информативную бета-компоненту, значительно растут как эффективность предлагаемого метода (при увеличении *n* от 1 до 50 прирост ROC AUC составил 26 %, а Recall@100 — 84 %), так и эффективность метода Фишера (ROC AUC — 67 %, Recall@100 — 353 %). Такой рост значений метрик качества свидетельствует о том, что дополнительная информация (большее число экспериментов) улучшает определение вершин активного модуля. Предлагаемый метод устанавливает активные модули преимущественно эффективнее, чем метод Фишера по обеим метрикам качества, особенно при *n* от 1 до 10. При $n = 50$ результаты были схожи для обоих методов.

При увеличении *n* в случае $\alpha = 0{,}8$, когда бета-равномерное распределение ближе к равномерному, незначительно растет как эффективность предлагаемого метода (при увеличении *n* от 1 до 50 прирост ROC AUC составил 10 %, Recall@100 — 42 %), так и эффективность метода Фишера (ROC AUC — 6 %, Recall@100 — 17 %). Предлагаемый метод определяет активные модули эффективнее, чем метод Фишера по обеим метрикам качества для всех *n*.

Для вариантов параметров $\alpha \in \{0{,}2;\ 0{,}8\}$ и $n \in \{1, 3, 10\}$ предлагаемый метод демонстрирует более высокие значения ROC AUC и Recall@100, чем метод Фишера. Однако при $\alpha = 0{,}2$ и $n = 50$ метод Фишера показал более высокое значение метрики ROC AUC, что указывает на то, что в этих условиях информации, полученной только из экспериментов, уже достаточно, а учет взаимодействий в графе становится избыточным. Дальнейшее повышение числа экспериментов *n* является нецелесообразным для данного набора параметров *G* и α в графе InWebIM.

Суммарно, предлагаемый метод демонстрирует более высокую или равную точность определения вершин активного модуля по сравнению с ранжированием вершин на основе метода Фишера, что подтверждается лучшими значениями ROC AUC (рис. 2, *a*–*d*) и Recall@100 (рис. 2, *e*–*h*) в большинстве симуляций с ростом значения α.

Предлагаемый метод может быть применен к различным однородным биологическим графам. В качестве примера были выбраны два независимых графа: граф нейронов в головном мозге человека BigBrain (177 584 вершины и 15 669 037 ребер) и генный граф вида живых организмов *Caenorhabditis elegans* (HS-LC) (4227 вершин и 39 484 ребра).

К обоим графам был применен предлагаемый метод. Приращения метрик ROC AUC и Recall@100 для предлагаемого метода были аналогичны приращениям этих метрик на графе InWebIM по сравнению с методом Фишера за исключением большого числа экспериментов ($n = 50$). При таком числе экспериментов и высокой информативности каждого отдельного эксперимента ($\alpha < 0{,}6$) метод Фишера несет достаточную информацию о принадлежности вершины к активному модулю. Информация о топологии графа в таком случае становится избыточной, что приводит к отсутствию прироста ROC AUC предлагаемого метода по сравнению с методом Фишера как в случае графа BigBrain (рис. 3), так и снижению качества ROC AUC в среднем на 6 % для генного графа (HS-LC) при $\alpha < 0{,}6$ (рис. 4). Recall@100 в среднем ухудшается на 8 % для графа BigBrain и на 10 % для графа (HS-LC) при $\alpha < 0{,}6$.

Валидация предлагаемого метода с использованием результатов биологических экспериментов

Генетические исследования шизофрении

Один из основных методов определения связи между ДНК и болезнью — GWAS. Результатом GWAS является множество локусов (отрезков ДНК, включающих от одного до нескольких генов) и их статистическую значимость.

1. Возникает задача приоритизации генов внутри локуса по отношению к болезни [16, 17].

2. Систематическое сравнение точности метода Фишера и предлагаемого метода проводилось на примере решения задачи из п. 1. Было использовано

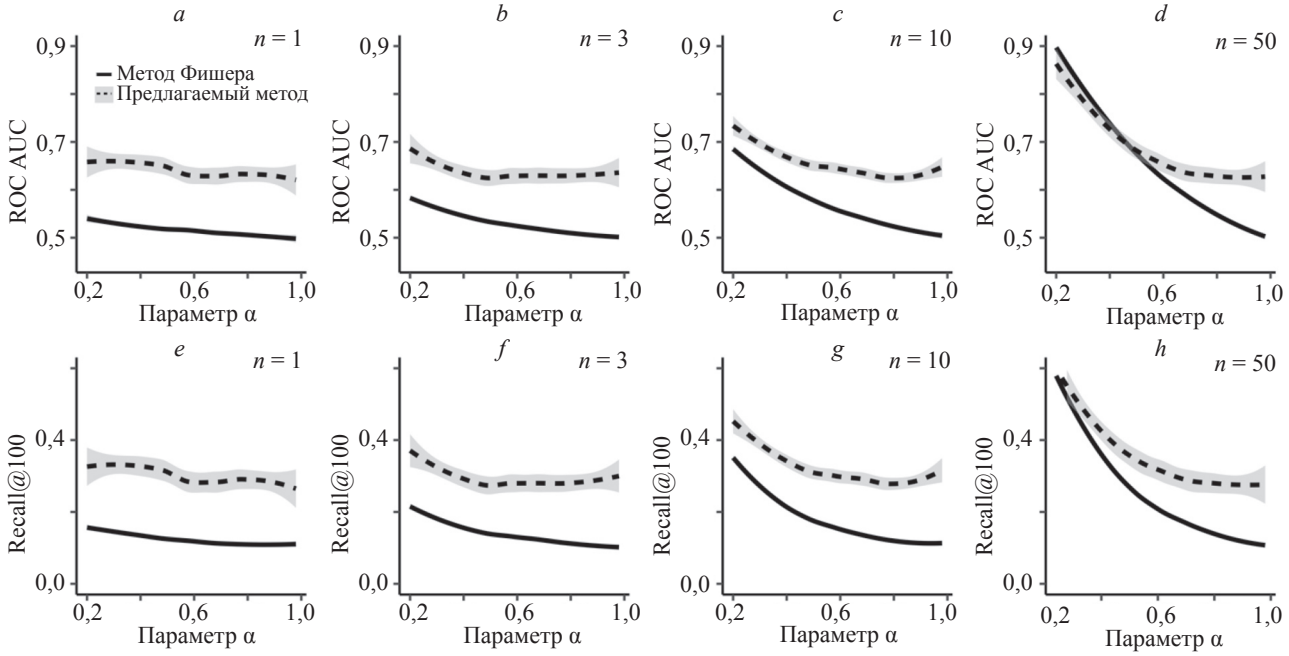

Рис. 2. Результаты симуляций для InWebIM. Сравнение метрик ROC AUC и Recall@100 в зависимости от параметра α бета-равномерного распределения и числа экспериментов. ROC AUC (*a*–*d*); Recall@100 (*e*–*h*)

Fig. 2. Simulation results for InWebIM. Comparison of ROC AUC and Recall@100 metrics vs. the parameter α of the beta-uniform distribution and the number of experiments. ROC AUC (*a*–*d*); Recall@100 (*e*–*h*)

GWAS по шизофрении [18]. 49 *p*-значений для каждого гена были получены методом, описанным в [19, 20] из GWAS по шизофрении. В качестве генов активного модуля были выбраны 105 известных генов шизофрении из работ [21–23]. В качестве генов вне активного модуля выбраны гены, расположенные на расстоянии 200 000 пар оснований от середины каждого гена активного модуля. Каждый локус включал не менее двух генов. Один эксперимент соответствовал одному из 49 полученных *p*-значений для каждого гена. Эксперимент выбирался случайно с помощью генератора псевдослучайных чисел.

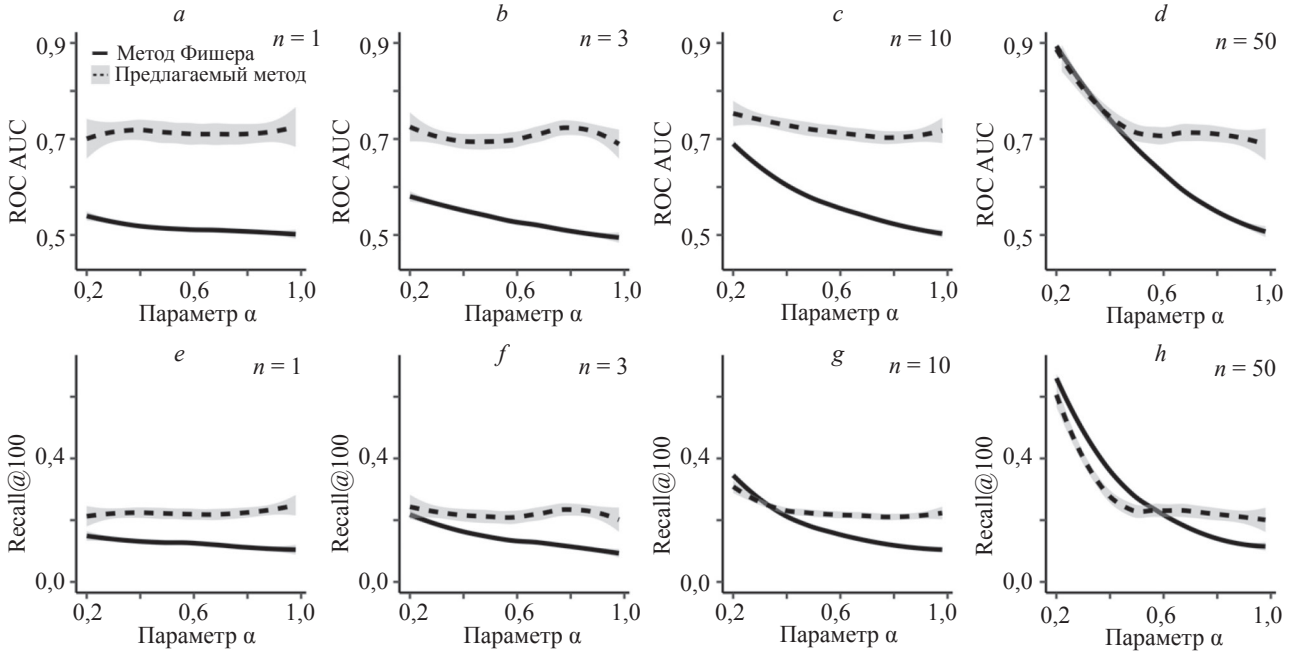

Рис. 3. Результаты симуляций для графа реконструкции мозга человека BigBrain. Сравнение метрик ROC AUC и Recall@100 в зависимости от параметра α бета-равномерного распределения и числа экспериментов. ROC AUC (*a*–*d*); Recall@100 (*e*–*h*)

Fig. 3. Simulation results for the BigBrain reconstruction network. Comparison of ROC AUC and Recall@100 metrics vs. the parameter α of the beta-uniform distribution and the number of experiments. ROC AUC (*a*–*d*); Recall@100 (*e*–*h*)

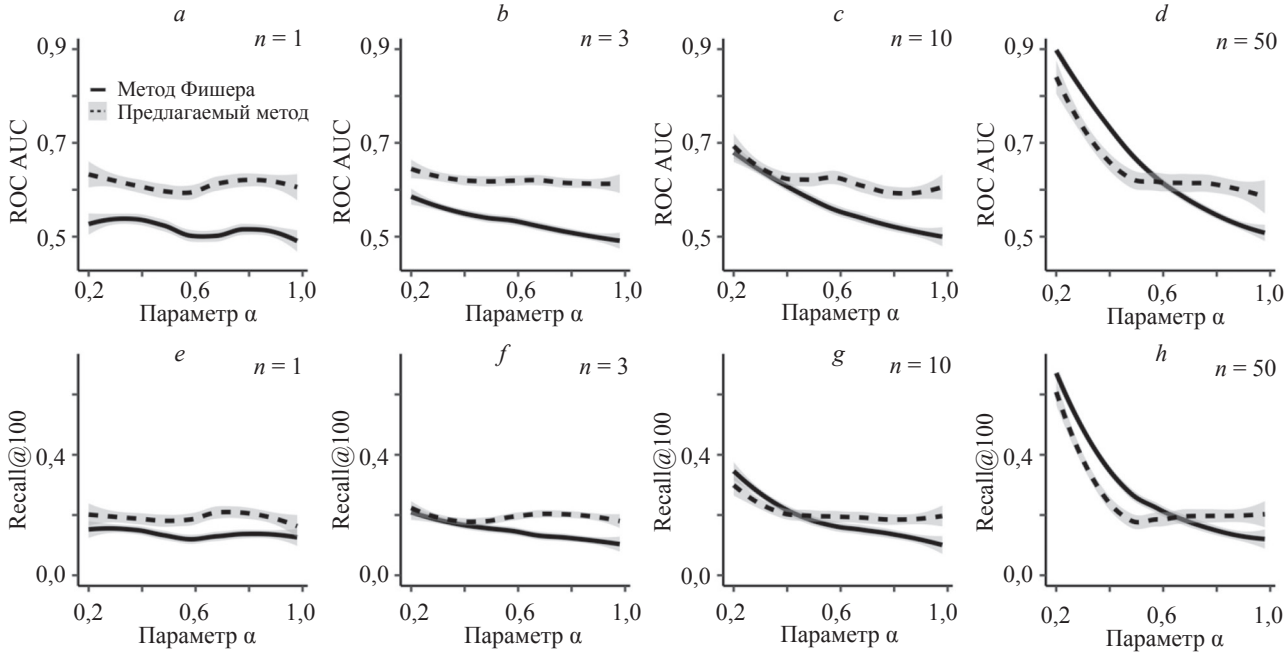

Рис. 4. Результаты симуляций для генного графа вида живых организмов Caenorhabditis elegans. Сравнение метрик ROC AUC и Recall@100 в зависимости от параметра α бета-равномерного распределения и числа экспериментов. ROC AUC (*a*–*d*); Recall@100 (*e*–*h*)

Fig. 4. Simulation results for the gene network of Caenorhabditis elegans. Comparison of ROC AUC and Recall@100 metrics vs. the parameter α of the beta-uniform distribution and the number of experiments. ROC AUC (*a*–*d*); Recall@100 (*e*–*h*)

3. Применен предлагаемый метод с учетом того, что результаты метода МКМЦ, нормализованы в диапазоне от нуля до единицы для каждого локуса независимо. В качестве биологического графа использован граф белок-белковых взаимодействий InWebIM, где каждому белку соответствовал один ген. Известными генами активного модуля для предлагаемого метода назначались 10 случайных генов из 105 известных. Остальные гены использовались для валидации предлагаемого метода. Этот метод и метод Фишера были применены 100 раз.

4. Точность метода Фишера и предлагаемого метода оценивалась с помощью метрик качества ROC AUC и Recall@200 (доля правильно определенных генов активного модуля в первых 200 вершинах).

5. Из рассмотрения коробчатых диаграмм следует, что предлагаемый метод в среднем показал результаты лучше, чем метод Фишера. ROC AUC предлагаемого метода для одного эксперимента был на 13 % выше, для трех экспериментов на 8 % выше, для 10 экспериментов на 5 % выше, для 49 экспериментов на 6 % выше (рис. 5, *a*). Recall@200 для предлагаемого метода был

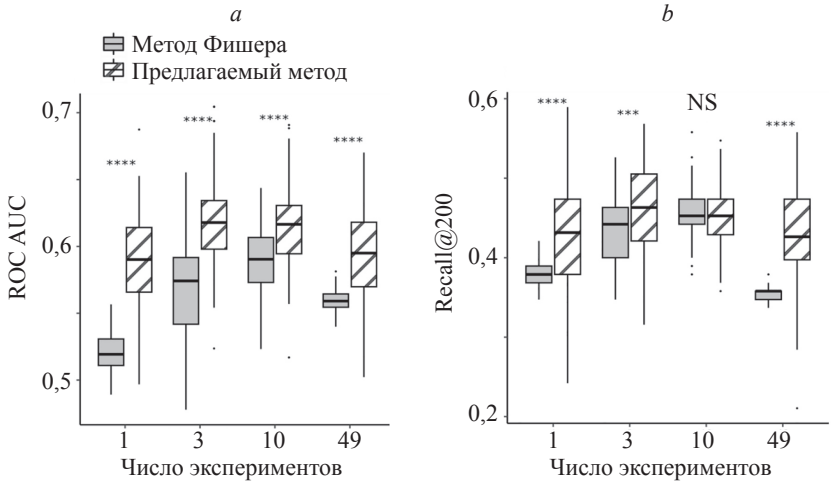

Рис. 5. Коробчатые диаграммы для предлагаемого метода и метода Фишера для GWAS по шизофрении. ROC AUC (*a*), Recall@200 (*b*); t-тест использован для сравнения метрик между методами (**** — *p*-значение < 0,0001; *** — *p*-значение < 0,001; NS — незначимо)

Fig. 5. Application of the proposed method and Fisher's method for GWAS on schizophrenia. ROC AUC (*a*), Recall@200 (*b*); the t-test was used to compare metrics between methods (**** — *p*-value < 0.0001; **** — *p*-value < 0.001; NS — not significant)

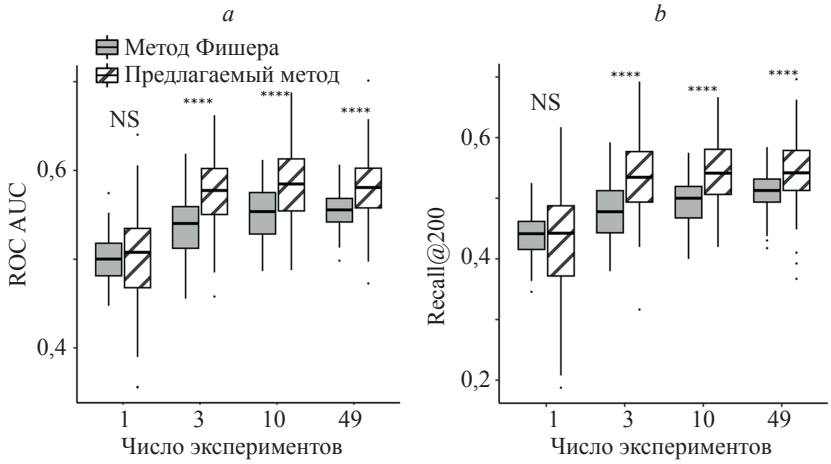

Рис. 6. Коробчатые диаграммы для предлагаемого метода и метода Фишера для данных GWAS по пристрастию к курению. ROC AUC (*a*), Recall@200 (*b*); t-тест использован для сравнения метрик между методами (**** — *p*-значение < 0,0001; NS — незначимо)

Fig. 6. Application of the proposed method and Fisher's method to GWAS data on smoking addiction. ROC AUC (*a*), Recall@200 (*b*); the t-test was used to compare metrics between methods (**** — *p*-value < 0.0001; NS — not significant)

лучше метода Фишера в среднем на 12 % для одного эксперимента, на 6 % для трех экспериментов и на 20 % для 49 экспериментов. В случае 10 экспериментов Recall@200 предлагаемого метода и метода Фишера не отличался (рис. 5, *b*).

Генетические исследования в Российской популяции

Состоялся релиз первой базы данных GWAS исследований в России (Биобанк России) [24]. Метод Фишера и предлагаемый метод были применены для такого фенотипа, как генетические риски пристрастия к курению из Биобанка России[1]. 175 известных гена для данного фенотипа были взяты из GWAS-каталога[2]. В качестве известных генов активного модуля для предлагаемого метода были выбраны 9 случайных генов из 175 известных. Остальные гены использовались для валидации предлагаемого метода. В качестве биологического графа применен граф белок-белковых взаимодействий InWebIM. ROC AUC для предлагаемого метода не отличался от метода Фишера для одного эксперимента. Recall@200 для предлагаемого метода был лучше метода Фишера в среднем на 11 % для трех экспериментов, на 9 % для 10 экспериментов и на 6 % для 49 экспериментов (рис. 6, *b*).

Генетические исследования кожной меланомы

В [25] описаны результаты исследования, в котором оценивалось влияние экспрессии каждого отдельного гена на выживаемость пациентов с кожной меланомой в двух независимых группах. К этим результатам были применены метод Фишера и предлагаемый метод. В качестве известных генов использованы 11 иммунных генов из [25]. Биологический граф построен из данных экспрессии генов методом семплирования байесовских сетей [26].

Для того чтобы оценить правильность полученного активного модуля, гены были ранжированы по предсказанным значениям принадлежности к активному модулю предлагаемым методом и методом Фишера независимо. Затем к первой 1000 генов из каждого списка независимо применен метод обогащения биологических путей [27]. Всего было выделено 62 биологических пути. Результаты статистической значимости для метода обогащения биологических путей на основе результатов предлагаемого метода и метода Фишера показаны на рис. 7. Каждая точка соответствует статистической значимости для метода обогащения биоло-

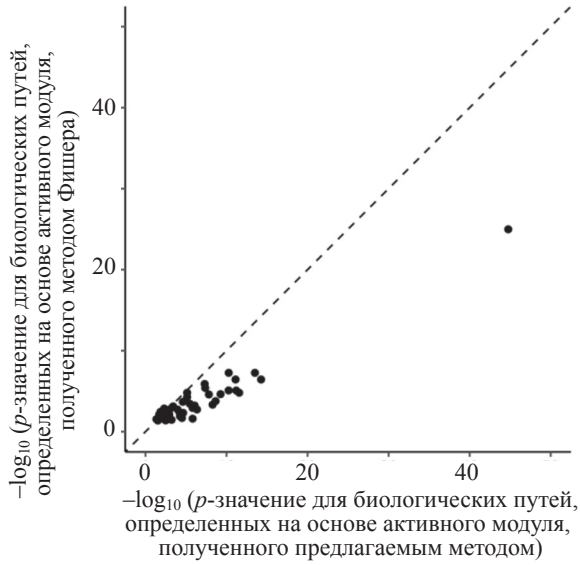

Рис. 7. Оценка качества полученного активного модуля кожной меланомы предлагаемым методом и методом Фишера

Fig. 7. The quality of the obtained active module of skin melanoma by the proposed method and the Fisher method

[1] Биобанк России [Электронный ресурс]. Режим доступа: https://biobankrus.almazovcentre.ru/pheno/SMNE, свободный. Яз. рус. (дата обращения: 15.12.2024).

[2] GWAS-catalog [Электронный ресурс]. Режим доступа: https://www.ebi.ac.uk/gwas/, свободный. Яз. англ. (дата обращения: 15.12.2024).

гических путей. Смещение точек вправо относительно диагональной кривой (рис. 7, пунктирная кривая), свидетельствует о том, что результаты метода обогащения биологических путей, полученные на основании результатов предлагаемого метода, статистически надежнее, чем на основании результатов, полученных методом Фишера.

Заключение

В работе предложен метод определения активных модулей в биологических графах с использованием многокомпонентных весов вершин. Этот метод основан на последовательном применении метода Фишера для интеграции данных нескольких биологических экспериментов, метода Монте-Карло по схеме марковских цепей для генерации вероятностных выборок связных подграфов, и машинного обучения — градиентного бустинга для учета топологии графа.

Экспериментальная оценка, выполненная на симулированных активных модулях графа белок-белковых взаимодействий InWebIM, а также на независимых биологических графах, таких как граф реконструкции мозга BigBrain и генный граф вида живых организмов — *Caenorhabditis elegans*, продемонстрировала значительное повышение точности предлагаемого метода по сравнению с методом Фишера, который не учитывает топологию графа. В частности, было показано, что точность определения активных модулей предлагаемым методом увеличивается с ростом числа учитываемых биологических экспериментов.

Валидация на реальных данных, полученных в генетических исследованиях шизофрении и курения, а также на экспрессионных данных пациентов с меланомой, подтверждает преимущества предлагаемого метода: он демонстрирует более высокое качество идентификации активных модулей по сравнению с методом Фишера.

Важным ограничением предложенного метода является предположение о независимости статистик из разных биологических экспериментов. Если между экспериментами существует сильная корреляция (например, одинаковые образцы или пересекающиеся данные), классический метод Фишера может переоценивать значимость итоговых результатов. В подобных случаях необходимы дополнительные корректировки или расширенные подходы к объединению статистик, учитывающие наличие корреляций между экспериментами.

Литература

1. Wang S., Wu R., Lu J., Jiang Y., Huang T., Cai Y.D. Protein-protein interaction networks as miners of biological discovery // Proteomics. 2022. V. 22. N 15-16. P. e2100190. https://doi.org/10.1002/pmic.202100190
2. Rao X., Dixon R.A. Co-expression networks for plant biology: why and how // Acta Biochimica et Biophysica Sinica. 2019. V. 51. N 10. P. 981–988. https://doi.org/10.1093/abbs/gmz080
3. Rawls K., Dougherty B.V., Papin J. Metabolic network reconstructions to predict drug targets and off-target effects // Methods in Molecular Biology. 2020. V. 2088. P. 315–330. https://doi.org/10.1007/978-1-0716-0159-4_14
4. Dittrich M.T., Klau G.W., Rosenwald A., Dandekar T., Müller T. Identifying functional modules in protein-protein interaction networks: an integrated exact approach // Bioinformatics. 2008. V. 24. N 13. P. i223–i231. https://doi.org/10.1093/bioinformatics/btn161
5. Zhu Q.M., Hsu Y.H.H., Lassen F.H., MacDonald B.T., Stead S., Malolepsza E., Kim A., Li T., Mizoguchi T., Schenone M., Guzman G., Tanenbaum B., Fornelos N., Carr S.A., Gupta R.M., Ellinor P.T., Lage K. Protein interaction networks in the vasculature prioritize genes and pathways underlying coronary artery disease // Communications Biology. 2024. V. 7. N 1. P. 87. https://doi.org/10.1038/s42003-023-05705-1
6. Nehme R., Pietiläinen O., Artomov M., Tegtmeyer M., Valakh V., Lehtonen L., Bell C., Singh T., Trehan A., Sherwood J. et. al. The 22q11.2 region regulates presynaptic gene-products linked to schizophrenia // Nature Communications. 2022. V. 13. N 1. P. 3690. https://doi.org/10.1038/s41467-022-31436-8
7. Alexeev N., Isomurodov J., Sukhov V., Korotkevich G., Sergushichev A. Markov chain Monte Carlo for active module identification problem // BMC Bioinformatics. 2020. V. 21. Suppl. 6. P. 261. https://doi.org/10.1186/s12859-020-03572-9
8. Усольцев Д.А., Молотков И.И., Артемов Н.Н., Сергушичев А.А., Шалыто А.А. Применение марковских цепей Монте-Карло и машинного обучения для поиска активного модуля в биологических графах // Научно-технический вестник информационных технологий, механики и оптики. 2024. Т. 24. № 6. С. 962–971. https://doi.org/10.17586/2226-1494-2024-24-6-962-971
9. Kim T.K., Park J.H. More about the basic assumptions of t-test: normality and sample size // Korean Journal of Anesthesiology. 2019. V. 72. N 4. P. 331–335. https://doi.org/10.4097/kja.d.18.00292

References

1. Wang S., Wu R., Lu J., Jiang Y., Huang T., Cai Y.D. Protein-protein interaction networks as miners of biological discovery. *Proteomics*, 2022, vol. 22, no. 15-16, P. e2100190. https://doi.org/10.1002/pmic.202100190
2. Rao X., Dixon R.A. Co-expression networks for plant biology: why and how. *Acta Biochimica et Biophysica Sinica*, 2019, vol. 51, no. 10, pp. 981–988. https://doi.org/10.1093/abbs/gmz080
3. Rawls K., Dougherty B.V., Papin J. Metabolic network reconstructions to predict drug targets and off-target effects. *Methods in Molecular Biology*, 2020, vol. 2088, pp. 315–330. https://doi.org/10.1007/978-1-0716-0159-4_14
4. Dittrich M.T., Klau G.W., Rosenwald A., Dandekar T., Müller T. Identifying functional modules in protein-protein interaction networks: an integrated exact approach. *Bioinformatics*, 2008, vol. 24, no. 13. pp. i223–i231. https://doi.org/10.1093/bioinformatics/btn161
5. Zhu Q.M., Hsu Y.H.H., Lassen F.H., MacDonald B.T., Stead S., Malolepsza E., Kim A., Li T., Mizoguchi T., Schenone M., Guzman G., Tanenbaum B., Fornelos N., Carr S.A., Gupta R.M., Ellinor P.T., Lage K. Protein interaction networks in the vasculature prioritize genes and pathways underlying coronary artery disease. *Communications Biology*, 2024, vol. 7, no. 1, pp. 87. https://doi.org/10.1038/s42003-023-05705-1
6. Nehme R., Pietiläinen O., Artomov M., Tegtmeyer M., Valakh V., Lehtonen L., Bell C., Singh T., Trehan A., Sherwood J. et. al. The 22q11.2 region regulates presynaptic gene-products linked to schizophrenia. *Nature Communications*, 2022, vol. 13, no. 1, pp. 3690. https://doi.org/10.1038/s41467-022-31436-8
7. Alexeev N., Isomurodov J., Sukhov V., Korotkevich G., Sergushichev A. Markov chain Monte Carlo for active module identification problem. *BMC Bioinformatics*, 2020, vol. 21, Suppl. 6, pp. 261. https://doi.org/10.1186/s12859-020-03572-9
8. Usoltsev D.A., Molotkov I.I., Artomov M.N., Sergushichev A.A., Shalyto A.A. Application of Markov chain Monte Carlo and machine learning for identifying active modules in biological graphs. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2024, vol. 24, no. 6, pp. 962–971. (in Russian). https://doi.org/10.17586/2226-1494-2024-24-6-962-971
9. Kim T.K., Park J.H. More about the basic assumptions of t-test: normality and sample size. *Korean Journal of Anesthesiology*, 2019, vol. 72, no. 4, pp. 331–335. https://doi.org/10.4097/kja.d.18.00292

10. Pounds S., Morris S.W. Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values // Bioinformatics. 2003. V. 19. N 10. P. 1236–1242. https://doi.org/10.1093/bioinformatics/btg148
11. Ham H., Park T. Combining p-values from various statistical methods for microbiome data // Frontiers in Microbiology. 2022. V. 13. P. 990870. https://doi.org/10.3389/fmicb.2022.990870
12. Li T., Wernersson R., Hansen R.B., Horn H., Mercer J., Slodkowicz G., Workman C.T., Rigina O., Rapacki K., Stærfeldt H.H., Brunak S., Jensen T.S., Lage K. A scored human protein-protein interaction network to catalyze genomic interpretation // Nature Methods. 2017. V. 14. N 1. P. 61–64. https://doi.org/10.1038/nmeth.4083
13. Rossi R., Ahmed N. The network data repository with interactive graph analytics and visualization // Proc. of the 29th AAAI Conference on Artificial Intelligence. 2015. V. 29. N 1. https://doi.org/10.1609/aaai.v29i1.9277
14. Amunts K., Lepage C., Borgeat L., Mohlberg H., Dickscheid T., Rousseau M.É., Bludau S., Bazin P.L., Lewis L.B., Oros-Peusquens A.M., Shah N.J., Lippert T., Zilles K., Evans A.C. BigBrain: an ultrahigh-resolution 3D human brain model // Science. 2013. V. 340. N 6139. P. 1472–1475. https://doi.org/10.1126/science.1235381
15. Cho A., Shin J., Hwang S., Kim C., Shim H., Kim H., Kim H., Lee I. WormNet v3: a network-assisted hypothesis-generating server for Caenorhabditis elegans // Nucleic Acids Research. 2014. V. 42. N W1. P. W76–W82. https://doi.org/10.1093/nar/gku367
16. Zhu Z., Zhang F., Hu H., Bakshi A., Robinson M.R., Powell J.E., Montgomery G.W., Goddard M.E., Wray N.R., Visscher P.M., Yang J. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets // Nature Genetics. 2016. V. 48. N 5. P. 481–487. https://doi.org/10.1038/ng.3538
17. Usoltsev D., Molotkov I., Artomov M. A meta-predictor for causal gene identification in GWAS overcomes limitations of existing computational approaches // American Society of Human Genetics (Complex Traits and Polygenic Disorders Poster Friday Session). 2024.
18. Pardiñas A.F., Holmans P., Pocklington A.J., Escott-Price V., Ripke S., Carrera N., Legge S.E., Bishop S., Cameron D., Hamshere M.L., et al. Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection // Nature Genetics. 2018. V. 50. N 3. P. 381–389. https://doi.org/10.1038/s41588-018-0059-2
19. Barbeira A.N., Dickinson S.P., Bonazzola R., Zheng J., Wheeler H.E., Torres J.M., Torstenson E.S., Shah K.P., Garcia T., Edwards T.L., Stahl E.A., Huckins L.M., Nicolae D.L., Cox N.J., Im H.K. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics // Nature Communications. 2018. V. 9. N 1. P. 1825. https://doi.org/10.1038/s41467-018-03621-1
20. Urbut S.M., Wang G., Carbonetto P., Stephens M. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions // Nature Genetics. 2019. V. 51. N 1. P. 187–195. https://doi.org/10.1038/s41588-018-0268-8
21. Kolosov N., Daly M.J., Artomov M. Prioritization of disease genes from GWAS using ensemble-based positive-unlabeled learning // European Journal of Human Genetics. 2021. V. 29. N 10. P. 1527–1535. https://doi.org/10.1038/s41431-021-00930-w
22. Lam M., Chen C-Y., Li Z., Martin A.R., Bryois J., Ma X., Gaspar H., Ikeda M., Benyamin B., Brown B.C. et al. Comparative genetic architectures of schizophrenia in East Asian and European populations // Nature Genetics. 2019. V. 51. N 12. P. 1670–1678. https://doi.org/10.1038/s41588-019-0512-x
23. Singh T., Poterba T., Curtis D., Akil H., Al Eissa M., Barchas J.D., Bass N., Bigdeli T.B., Breen G., Bromet E.J. et al. Rare coding variants in ten genes confer substantial risk for schizophrenia // Nature. 2022. V. 604. N 7906. P. 509–516. https://doi.org/10.1038/s41586-022-04556-w
24. Usoltsev D., Kolosov N., Rotar O., Loboda A., Boyarinova M., Moguchaya E., Kolesova E., Erina A., Tolkunova K., Rezapova V., Molotkov I. et al. Complex trait susceptibilities and population diversity in a sample of 4,145 Russians // Nature Communications. 2024. V. 15. N 1. P. 6212. https://doi.org/10.1038/s41467-024-50304-1
25. Usoltsev D., Njauw C.N., Ji Z., Kumar R., Sergushichev A., Zhang S., Shlyakhto E., Daly M.J., Artomov M., Tsao H. Analysis of variants induced by combined ex vivo irradiation and in vivo tumorigenesis

10. Pounds S., Morris S.W. Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics*, 2003, vol. 19, no. 10, pp. 1236–1242. https://doi.org/10.1093/bioinformatics/btg148
11. Ham H., Park T. Combining p-values from various statistical methods for microbiome data. *Frontiers in Microbiology*, 2022, vol. 13, pp. 990870. https://doi.org/10.3389/fmicb.2022.990870
12. Li T., Wernersson R., Hansen R.B., Horn H., Mercer J., Slodkowicz G., Workman C.T., Rigina O., Rapacki K., Stærfeldt H.H., Brunak S., Jensen T.S., Lage K. A scored human protein-protein interaction network to catalyze genomic interpretation. *Nature Methods*, 2017, vol. 14, no. 1, pp. 61–64. https://doi.org/10.1038/nmeth.4083
13. Rossi R., Ahmed N. The network data repository with interactive graph analytics and visualization. *Proc. of the 29th AAAI Conference on Artificial Intelligence*, 2015, vol. 29, no. 1. https://doi.org/10.1609/aaai.v29i1.9277
14. Amunts K., Lepage C., Borgeat L., Mohlberg H., Dickscheid T., Rousseau M.É., Bludau S., Bazin P.L., Lewis L.B., Oros-Peusquens A.M., Shah N.J., Lippert T., Zilles K., Evans A.C. BigBrain: an ultrahigh-resolution 3D human brain model. *Science*, 2013, vol. 340, no. 6139, pp. 1472–1475. https://doi.org/10.1126/science.1235381
15. Cho A., Shin J., Hwang S., Kim C., Shim H., Kim H., Kim H., Lee I. WormNet v3: a network-assisted hypothesis-generating server for Caenorhabditis elegans. *Nucleic Acids Research*, 2014, vol. 42, no. W1, pp. W76–W82. https://doi.org/10.1093/nar/gku367
16. Zhu Z., Zhang F., Hu H., Bakshi A., Robinson M.R., Powell J.E., Montgomery G.W., Goddard M.E., Wray N.R., Visscher P.M., Yang J. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nature Genetics*, 2016, vol. 48, no. 5, pp. 481–487. https://doi.org/10.1038/ng.3538
17. Usoltsev D., Molotkov I., Artomov M. A meta-predictor for causal gene identification in GWAS overcomes limitations of existing computational approaches. *American Society of Human Genetics (Complex Traits and Polygenic Disorders Poster Friday Session)*, 2024.
18. Pardiñas A.F., Holmans P., Pocklington A.J., Escott-Price V., Ripke S., Carrera N., Legge S.E., Bishop S., Cameron D., Hamshere M.L., et. al. Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nature Genetics*, 2018, vol. 50, no. 3, pp. 381–389. https://doi.org/10.1038/s41588-018-0059-2
19. Barbeira A.N., Dickinson S.P., Bonazzola R., Zheng J., Wheeler H.E., Torres J.M., Torstenson E.S., Shah K.P., Garcia T., Edwards T.L., Stahl E.A., Huckins L.M., Nicolae D.L., Cox N.J., Im H.K. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nature Communications*, 2018, vol. 9, no. 1, pp. 1825. https://doi.org/10.1038/s41467-018-03621-1
20. Urbut S.M., Wang G., Carbonetto P., Stephens M. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nature Genetics*, 2019, vol. 51, no. 1, pp. 187–195. https://doi.org/10.1038/s41588-018-0268-8
21. Kolosov N., Daly M.J., Artomov M. Prioritization of disease genes from GWAS using ensemble-based positive-unlabeled learning. *European Journal of Human Genetics*, 2021, vol. 29, no. 10, pp. 1527–1535. https://doi.org/10.1038/s41431-021-00930-w
22. Lam M., Chen C-Y., Li Z., Martin A.R., Bryois J., Ma X., Gaspar H., Ikeda M., Benyamin B., Brown B.C. et al. Comparative genetic architectures of schizophrenia in East Asian and European populations. *Nature Genetics*, 2019, vol. 51, no. 12, pp. 1670–1678. https://doi.org/10.1038/s41588-019-0512-x
23. Singh T., Poterba T., Curtis D., Akil H., Al Eissa M., Barchas J.D., Bass N., Bigdeli T.B., Breen G., Bromet E.J. et al. Rare coding variants in ten genes confer substantial risk for schizophrenia. *Nature*, 2022, vol. 604, no. 7906, pp. 509–516. https://doi.org/10.1038/s41586-022-04556-w
24. Usoltsev D., Kolosov N., Rotar O., Loboda A., Boyarinova M., Moguchaya E., Kolesova E., Erina A., Tolkunova K., Rezapova V., Molotkov I. et al. Complex trait susceptibilities and population diversity in a sample of 4,145 Russians. *Nature Communications*, 2024, vol. 15, no. 1, pp. 6212. https://doi.org/10.1038/s41467-024-50304-1
25. Usoltsev D., Njauw C.N., Ji Z., Kumar R., Sergushichev A., Zhang S., Shlyakhto E., Daly M.J., Artomov M., Tsao H. Analysis of variants induced by combined ex vivo irradiation and in vivo tumorigenesis

suggests a role for the ZNF831 p.R1393Q variant in cutaneous melanoma development // Journal of Investigative Dermatology. 2024. In Press, Corrected Proof. https://doi.org/10.1016/j.jid.2024.08.042
26. Лобода А.А. Метод графовой кластеризации для совместного анализа данных генотипирования и экспрессии генов: диссертация на соискание ученой степени кандидата технических наук. СПб., 2022, 232 с.
27. Subramanian A., Tamayo P., Mootha V.K., Mukherjee S., Ebert B.L., Gillette M.A., Paulovich A., Pomeroy S.L., Golub T.R., Lander E.S., Mesirov J.P. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles // Proc. of the National Academy of Sciences of the United States of America. 2005. V. 102. N 43. P. 15545–15550. https://doi.org/10.1073/pnas.0506580102

suggests a role for the ZNF831 p.R1393Q variant in cutaneous melanoma development. *Journal of Investigative Dermatology*, 2024, in Press, corrected proof. https://doi.org/10.1016/j.jid.2024.08.042
26. Loboda A.A. A method of graphical clustering for joint analysis of genotyping and expression data. *Dissertation for the degree of candidate of technical sciences*. St. Petersburg, 2022, 232 p. (in Russian)
27. Subramanian A., Tamayo P., Mootha V.K., Mukherjee S., Ebert B.L., Gillette M.A., Paulovich A., Pomeroy S.L., Golub T.R., Lander E.S., Mesirov J.P. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. of the National Academy of Sciences of the United States of America*, 2005, vol. 102, no. 43, pp. 15545–15550. https://doi.org/10.1073/pnas.0506580102

Авторы

Усольцев Дмитрий Андреевич — старший научный сотрудник, Институт геномной медицины, Детская больница Нейшенвайд, Колумбус, 43205, США; аспирант, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 57279360300, https://orcid.org/0000-0001-8072-310X, dusoltsev.27@gmail.com

Молотков Иван Игоревич — старший научный сотрудник, Институт геномной медицины, Детская больница Нейшенвайд, Колумбус, 43205, США; аспирант, Медицинский колледж Университета штата Огайо, Колумбус, 43210, США, SC 58651494600, https://orcid.org/0009-0008-3566-0160, ivan.molotkov@nationwidechildrens.org

Артемов Никита Николаевич — кандидат химических наук, доцент, главный исследователь, Институт геномной медицины, Детская больница Нейшенвайд, Колумбус, 43205, США; профессор педиатрии, Медицинский колледж Университета штата Огайо, Колумбус, 43210, США, SC 36542095500, https://orcid.org/0000-0001-5282-8764, mykyta.artomov@nationwidechildrens.org

Сергушичев Алексей Александрович — кандидат технических наук, доцент, Университет Вашингтона в Сент-Луисе, Сент-Луис, 63110, США, SC 55772694000, https://orcid.org/0000-0003-1159-7220, asergushichev@wustl.edu

Шалыто Анатолий Абрамович — доктор технических наук, профессор, главный научный сотрудник, профессор, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 56131789500, https://orcid.org/0000-0002-2723-2077, anatoly.shalyto@gmail.com

Authors

Dmitrii A. Usoltsev — Senior Researcher, Institute for Genomic Medicine, Nationwide Children's Hospital, Columbus, 43205, USA; PhD Student, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 57279360300, https://orcid.org/0000-0001-8072-310X, dusoltsev.27@gmail.com

Ivan I. Molotkov — Senior Researcher, Institute for Genomic Medicine, Nationwide Children's Hospital, Columbus, 43205, USA; PhD Student, The Ohio State University College of Medicine, Columbus, 43210, USA, SC 58651494600, https://orcid.org/0009-0008-3566-0160, ivan.molotkov@nationwidechildrens.org

Mykyta N. Artomov — PhD (Chemistry), Associate Professor, Chief Researcher, Institute for Genomic Medicine, Nationwide Children's Hospital, Columbus, 43205, USA; Professor of Pediatrics, The Ohio State University College of Medicine, Columbus, 43210, USA, SC 36542095500, https://orcid.org/0000-0001-5282-8764, mykyta.artomov@nationwidechildrens.org

Alexey A. Sergushichev — PhD, Associate Professor, Washington University School of Medicine in St. Louis, St. Louis, 63110, USA, SC 55772694000, https://orcid.org/0000-0003-1159-7220, asergushichev@wustl.edu

Anatoly A. Shalyto — D.Sc., Chief Researcher, Full Professor, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 56131789500, https://orcid.org/0000-0002-2723-2077, anatoly.shalyto@gmail.com

Статья поступила в редакцию 09.02.2025
Одобрена после рецензирования 14.04.2025
Принята к печати 26.05.2025

Received 09.02.2025
Approved after reviewing 14.04.2025
Accepted 26.05.2025