НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
ноябрь–декабрь 2024 Том 24 № 6 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
November–December 2024 Vol. 24 No 6 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

doi: 10.17586/2226-1494-2024-24-6-962-971
УДК 004.021

Применение марковских цепей Монте-Карло и машинного обучения для поиска активного модуля в биологических графах

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

[1,5] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация
[1,2,3] Институт геномной медицины, Детская больница Нейшенвайд, Колумбус, 43205, США
[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. В биологии информация о взаимодействии изучаемых белков или генов может быть представлена в виде биологического графа. Связный подграф, вершины которого выполняют общую биологическую функцию, называется активным модулем. Марковская цепь Монте-Карло (MCMC) — эффективный алгоритм для идентификации активного модуля в биологических графах. В контексте белок-белковых взаимодействий точное нахождение активного модуля позволяет определить, какое нарушение белковой функции приводит к возникновению определенных изменений (например, болезни) в биологической системе (клетке/организме). Показано, что применение MCMC совместно с обучением моделей, учитывающих топологию графа, обеспечивает более высокую точность определения активного модуля. **Метод.** В работе независимо используется граф белок-белковых взаимодействий (InWebIM) и сеть функциональных ассоциаций между генами GeneMANIA для обучения модели и сравнения с известным методом на основе MCMC. В качестве методов поиска активного модуля использовалась комбинация из MCMC и метода машинного обучения — градиентного бустинга — xgboost. **Основные результаты.** Совместное применение метода на основе MCMC и xgboost повышает точность нахождения активного модуля по сравнению с методом на основе MCMC на симулированных данных. **Обсуждение.** Повышение точности поиска активного модуля имеет важное значение для исследования биологических механизмов заболеваний и обнаружения отдельных белков, функционально связанных с возникновением заболеваний.

Ключевые слова

графы, машинное обучение, белковые сети, MCMC, активный модуль

Ссылка для цитирования: 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. doi: 10.17586/2226-1494-2024-24-6-962-971

Application of Markov chain Monte Carlo and machine learning for identifying active modules in biological graphs

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

[1,5] ITMO University, Saint Petersburg, 197101, Russian Federation
[1,2,3] Institute for Genomic Medicine, Nationwide Chldren's Hospital, Columbus, 43205, USA
[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

In biology, information about interactions between the proteins or genes under study can be represented as a biological graph. A connected subgraph, whose vertices perform a common biological function, is called an active module. The Markov Chain Monte Carlo (MCMC) algorithm is an effective method for identifying active modules in biological graphs. In the context of protein-protein interactions, accurately identifying the active module allows for determining which protein function disruption leads to certain changes (e.g., diseases) in a biological system (cell/organism). This study demonstrates that applying MCMC in combination with models (that take graph topology into account) provides higher accuracy in identifying the active module. This study independently utilizes a protein-protein interaction graph (InWebIM) and the GeneMANIA functional association network for training the model and comparing it with the known MCMC-based method. To search for the active module, a combination of MCMC and a machine learning method, gradient boosting (xgboost), was employed. The combined use of the MCMC-based method and gradient boosting improves the accuracy of active module identification compared to the MCMC-based method alone on simulated data. Improving the accuracy of active module identification is crucial for studying the biological mechanisms of diseases and discovering individual proteins functionally associated with the development of diseases.

Keywords

graphs, machine learning, protein networks, MCMC, active module

For citation: 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). doi: 10.17586/2226-1494-2024-24-6-962-971

Введение

Представление биологической информации в виде графов является широко используемым методом для изучения различных типов взаимодействий, таких как ко-экспрессия генов и прямые взаимодействия между белками при построении функциональных белковых комплексов [1]. Например, белковый комплекс, который образует белок Bap1 совместно с 10 другими белками, обеспечивает функции восстановления дезоксирибонуклеиновой кислоты, участвует в клеточном цикле и дифференцировке клеток [2]. Нарушение работы Bap1-комплекса значительно повышает риски возникновения рака кожи [3].

Наличие прямых белок-белковых взаимодействий изучается экспериментально с использованием метода ко-иммунопреципитации [4] с последующей масс-спектрометрией [5]. С помощью антител, специфичных к белку-мишени, происходит осаждение (снятие гидратной оболочки и заряда), при этом также осаждаются белки, которые физически взаимодействуют с белком-мишенью. Последующая масс-спектрометрия позволяет распознать взаимодействующие белки [6].

Агрегация результатов таких экспериментов позволила создать каталог парных белок-белковых взаимодействий — сети InWebIM [7]. Он может быть представлен в виде биологического графа — связной структуры, в которой вершинами являются белки, а ребрами — взаимодействия между парами белков в клетке, известные из экспериментов по ко-иммунопреципитации. Граф состоит из 17 585 вершин и 657 411 ребер.

Внутри полного графа InWebIM можно выделить связанные подграфы, которые будут соответствовать белковым комплексам, вовлеченным в регуляцию отдельных биологических процессов. Такие связанные подграфы называют активными модулями (рис. 1).

В рассмотренном примере функционал белка Bap1 и взаимодействующих с ним белков известен: влияние на риск возникновения рака кожи. Однако для биологических состояний (болезней), этиология которых малоизучена, нахождение активного модуля, связанного с изучаемым состоянием — важная задача для понимания механизма возникновения такого состояния биологической системы. В применении к болезням — идентификация белковых активных модулей позволяет понять причины возникновения заболевания и определить возможные терапевтические мишени [8, 9].

В работах [10, 11] предложено несколько подходов к поиску активного модуля. Они включают следующие основные шаги: сбор исходных данных биологического эксперимента, определение весов вершин в графах и алгоритмический поиск активного модуля. Далее в работе каждый из этих шагов будет описан более подробно.

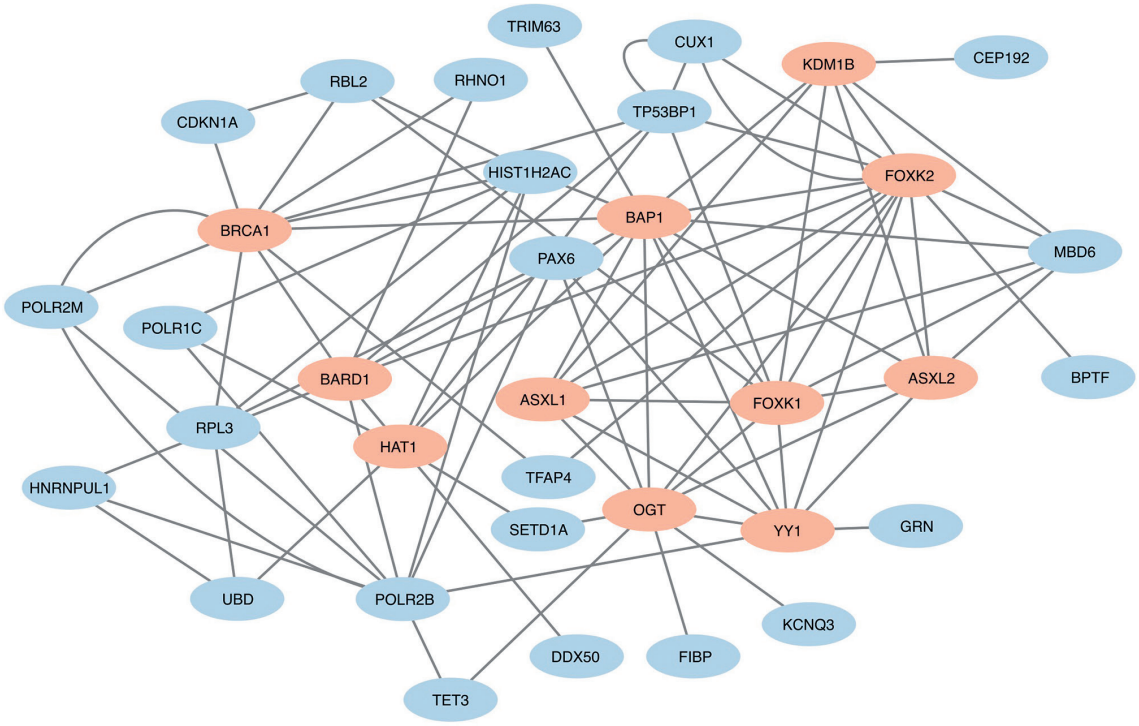

Рис. 1. Пример активного модуля. Bap1-комплекс, участвующий в восстановлении дезоксирибонуклеиновой кислоты. Красным цветом выделены белки, образующие Bap1-комплекс

Fig. 1. Example of an active module. The Bap1 complex involved in deoxyribonucleic acid repair. Proteins forming the Bap1 complex are shown in red

Во-первых, согласно центральной догме молекулярной биологии, каждому гену может быть поставлен в соответствие только один белок [12]. Любые изменения в генах транслируются в белки и приводят к формированию биологического состояния. Таким образом, множество белков является отображением множества генов. Поскольку набор генов в каждой клетке одинаков, а число молекул каждого вида белка вариабельно (включая ноль), эксперименты по выявлению генетических причин заболеваний целесообразно проводить на генах. Потому для получения первичной информации о функциональной вовлеченности белков в развитие болезни проводятся эксперименты по изучению дифференциальной экспрессии соответствующих им генов или сравнение мутационной нагрузки между группами пациентов с болезнью и контрольной группы [13, 14].

Во-вторых, в ходе анализа статистический тест (например, t-тест [15]) между группами проводится для каждого гена независимо. На выходе из эксперимента получается список генов с соответствующим *p*-значением. Распределение *p*-значений представляет собой сумму двух распределений — бета и равномерного [16]. Для бета-распределения, определенного на интервале [0, 1], его плотность вычисляется на основе соотношения:

$$f(x) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-\beta)^{\beta-1},\ 0 \le x \le 1;\ \alpha, \beta > 0, \quad (1)$$

где α и β — параметры формы бета-распределения; $B(\alpha, \beta)$ — бета-функция.

Бета-распределение (1) описывает гены, относящиеся к биологическому сигналу. Остальные гены принадлежат к статистическому шуму и могут быть описаны равномерным распределением $X \sim U[0, 1]$, где X — случайная величина, $U[0, 1]$ — равномерное распределение на интервале от 0 до 1.

Результирующее распределение называется бета-равномерным β(α, 1) и определяется плотностью, вычисляемой на основе соотношения:

$$f(x) = \lambda + (1-\lambda)\alpha x^{\alpha-1},\ 0 \le x, \lambda, \alpha \le 1, \quad (2)$$

где λ — вес равномерной компоненты; α — параметр формы бета-компоненты [17].

Так как *p*-значения, соответствующие биологическому сигналу, близки к нулю, ограничим параметр α сверху единицей.

Поскольку множество белков является отображением множества генов, то экспериментальные *p*-значения для генов могут быть присвоены соответствующим белкам. Далее белкам в биологическом графе посредством аппроксимации экспериментальных *p*-значений с помощью бета-равномерного распределения может быть присвоен соответствующий вес. Параметры бета-равномерного распределения (2) могут подбираться путем максимизации логарифма функции правдоподобия:

$$\log L(\alpha, \lambda | p_1, p_2, \ldots, p_n) = \sum_{i=1}^{n} \log(\lambda + (1-\lambda)\alpha p_i^{\alpha-1}), \quad (3)$$

где L — функция правдоподобия; p_i — *p*-значение для *i*-го белка.

Таким образом, имея граф со взвешенными вершинами, можно решить задачу поиска активного модуля для выявления белков, наиболее существенно влияющих на риски возникновения заболевания на основе нарушения белок-белковых взаимодействий.

Зная веса вершин, поиск активного модуля можно осуществить точным методом на основе подграфа максимального веса (Maximum-Weight Connected Subgraph, MWCS) [18] или вероятностно с помощью метода на основе марковской цепи Монте-Карло (Markov chain Monte Carlo, MCMC) [17].

Следует отметить, что полный список белков, влияющих на формирование биологического состояния обычно неизвестен. Таким образом, задача поиска активного модуля сводится к приоритизации — ранжированию белков по их вероятности включения в активный модуль. В этих условиях вероятностный метод на основе MCMC оказывается более предпочтительным по сравнению с методом на основе MWCS.

Часто для конкретного биологического состояния известны лишь некоторые белки активного модуля [19]. Это является основанием для предположения о том, что совместное использование метода на основе MCMC и информации об известных белках активного модуля, а также топологии графа InWebIM, может существенно улучшить определение ранее неизвестных белков этого модуля.

В ситуации, когда не существует точного решения для задачи определения активного модуля, оптимальным вариантом является симуляция активных модулей с заданными параметрами в реальной белок-белковой сети.

В настоящей работе выполнено моделирование активных модулей и известных белков в них на подграфах с числом вершин $N = 1000$ в белок-белковой сети InWebIM. Предложен метод поиска активных модулей, основанный на модели градиентного бустинга (xgboost), который учитывает метод на основе MCMC, а также расстояния от каждой вершины сети до известных белков, входящих в модуль. Эффективность предлагаемого метода продемонстрирована в сравнении с методом на основе MCMC. Кроме того, показана применимость и эффективность рассматриваемого метода на другом биологическом графе, независимом от графа InWebIM.

Предлагаемый метод

Этапы постановки математической задачи.

1. **Определение активных модулей.** В отсутствии достоверно известных активных модулей для большинства биологических состояний проводится симуляция активных модулей с G вершинами в сети InWebIM посредством поиска в ширину из одной случайной вершины. Для повышения скорости вычислений эта сеть сэмплируется на связанные подграфы — упрощенные белок-белковые сети. Число вершин в каждом подграфе $N = 1000$. Каждый подграф формируется поиском в ширину из одной случайной вершины.
2. **Определение признаков, характеризующих вершины сети.** M процентов от всех вершин активного модуля случайным образом выбираются как множество известных белков. Определяются расстояния от каждого белка упрощенной белок-белковые сети до трех ближайших известных белков. Определяются вероятности вхождения вершин в активный модуль с помощью метода на основе MCMC.
3. **Тренировка и тестирование модели.** Тренируется модель на основе машинного обучения, учитывающая расстояния до известных белков и результаты метода на основе MCMC. Модель определяет вероятность вхождения каждого белка упрощенной белок-белковой сети в активный модуль.

Модель тестируется с использованием таких метрик качества классификации белков по принадлежности активному модулю, как «площадь под кривой ошибок» (Receiver Operating Characteristic Area Under the Curve — ROC AUC — величина, варьируемая в диапазоне от нуля до единицы) и чувствительность (Recall@100 — доля правильно определенных белков активного модуля в первой сотне вершин, ранжированных по уменьшению предсказанной с помощью модели вероятности вхождения в активный модуль — величина, варьируемая в диапазоне от нуля до единицы).

ROC AUC определяет общую способность модели отличать белки, принадлежащие активному модулю, от тех, которые к нему не относятся. Recall@100 позволяет оценить эффективность приоритизации и определения наиболее важных белков активного модуля. Результаты тестирования модели с использованием ROC AUC и Recall@100 сравниваются с результатами тестирования по этим метрикам метода на основе MCMC.

Этапы решения задачи.

Этап 1. Симуляция 100 различных сетей с известными активными модулями.

1.1. Для получения сетей с топологией, встречающейся в реальных данных, выбираются случайные подграфы сети белок-белковых взаимодействий с числом вершин равным 1000.

1.2. В каждом подграфе белок-белковых взаимодействий из пункта 1.1 с помощью генератора псевдослучайных чисел выбирается вершина, от которой с помощью поиска в ширину определяется связный подграф с числом вершин, равным G. G равняется 5 или 10 % от N.

1.3. 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.

Этап 2. Вычисление признаков для модели, предсказывающей принадлежность каждой вершины активному модулю.

2.1. Три признака, описывающие каждый белок подграфа из этапа 1, определяются как расстояния до трех ближайших известных белков активного модуля.

2.2. Четвертый признак — вероятность вхождения белка в активный модуль, определяется по p-значениям из пункта 1.3. Для каждого белка вероятность

вычисляется посредством аппроксимации *p*-значений с помощью бета-равномерного распределения с параметрами, определенными по формуле (3). Это равносильно тому, что соответствующий белок относится к активному модулю. Далее 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 набирается выборка размера 100 с помощью алгоритма MCMC mcmcRanking (v0.1.0) [20], используя 10 000 симуляций MCMC. После этого для каждого белка определяется эмпирическая вероятность вхождения в активный модуль — доля подграфов из выборки, которые включают в себя этот белок.

Этап 3. Обучение модели xgboost для предсказания принадлежности вершины активному модулю.

Используя полученные признаки, обучается модель xgboost [21]. Она реализована в R-библиотеке xgboost (v1.5.0.2) [22]. Максимальная глубина каждого дерева в модели равняется трем. Параметр скорости обучения — 0,1. Число итераций обучения равно трем. 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 – ноль.

Эффективность предложенного метода

Симуляции проведены для модулей с количеством вершин $G = 100$ и $G = 50$. Выполнена оценка эффективности метода на основе MCMC и предлагаемого метода с использованием модели xgboost, с помощью метрик ROC AUC и Recall@100. Метрики определялись для каждого подграфа сети белок-белковых взаимодействий, после этого вычислялось среднее значение каждой метрики из выборки в 50 подграфов.

ROC AUC в зависимости от α бета-равномерного распределения при числе вершин активного модуля $G = 100$. Показано, что точность определения белков в активном модуле отражает ожидаемое поведение бета-равномерного распределения. При увеличении параметра α наблюдалось уменьшение ROC AUC для метода на основе MCMC, так как при стремлении параметра a к единице бета-распределение стремится к равномерному и веса активного модуля становятся мало отличимыми от равномерного распределения (рис. 2, *a*, сплошная кривая).

Добавление информации о расстоянии до трех ближайших известных белков из активного модуля значительно увеличивает ROC AUC. Рост ROC AUC для предлагаемого метода с использованием модели xgboost по сравнению с методом на основе MCMC составил от 6,3 % для параметра $\alpha = 0{,}2$ и до 23,1 % для параметра $\alpha = 0{,}2$ и при $M = 20$ % известных белков для каждого значения параметра α. Заметим, что разница в точности предлагаемого метода и метода на основе MCMC достигает максимума при более пологом бета-распределении (параметр $\alpha > 0{,}6$) (рис. 2, *a*, пунктирная кривая).

Recall@100 в зависимости от α бета-равномерного распределения при числе вершин активного модуля $G = 100$. Метрика Recall@100 показала согласованность с метрикой ROC AUC. Увеличение Recall@100 для предлагаемого метода с использованием модели xgboost по сравнению с методом на основе MCMC составил от 6,3 % для параметра $\alpha = 0{,}2$ и до 35,8 % для параметра $\alpha = 0{,}8$ и при $M = 20$ % известных белков для каждого значения параметра α (рис. 2, *b*).

ROC AUC в зависимости от процента известных белков в активном модуле при $\alpha = 0{,}5$ и числа вершин активного модуля $G = 100$. Был зафиксирован параметр $\alpha = 0{,}5$ и изменен процент известных белков M в активном модуле на интервале [1, 50]. Из рис. 2, *c* (пунктирная кривая) следует, что даже при наличии одного известного белка ($M = 1$ %, $G = 100$) для предлагаемого метода имеет место увеличение ROC AUC на 8,6 %. Увеличение процента известных белков до 37 % привело к повышению ROC AUC на 16,3 %. Отметим, что ROC AUC метода на основе MCMC не менялся, так как он не зависит от M.

Recall@100 в зависимости от процента известных белков в активном модуле при $\alpha = 0{,}5$ и числа вершин активного модуля $G = 100$. Увеличение Recall@100 для предлагаемого метода по сравнению с методом на основе MCMC составил от 13 % для $M = 1$ % и до 24,3 % для $M = 37$ % (рис. 2, *d*).

ROC AUC в зависимости от параметра α бета-равномерного распределения при числе вершин активного модуля $G = 50$. Число белков в активном модуле было уменьшено до $G = 50$. Повышение ROC AUC для предлагаемого метода по сравнению с методом на основе MCMC составил от 6,2 % для параметра $\alpha = 0{,}2$ и до 29,3 % для параметра $\alpha = 0{,}2$ и при $M = 20$ % (рис. 2, *e*).

Recall@100 в зависимости от параметра α бета-равномерного распределения при числе вершин активного модуля $G = 50$. Увеличение Recall@100 для предлагаемого метода по сравнению с методом на основе MCMC составил от 5 % для параметра $\alpha = 0{,}2$ и до 78,9 % для параметра $\alpha = 0{,}8$ и при $M = 20$ % (рис. 2, *f*).

ROC AUC в зависимости от процента известных белков в активном модуле при $\alpha = 0{,}5$ и числа вершин активного модуля $G = 50$. Был зафиксирован параметр $\alpha = 0{,}5$ и изменен процент известных белков M в активном модуле на интервале [2, 50]. Рост значения ROC AUC для предлагаемого метода по сравнению с методом на основе MCMC составил от 9,4 % для $M = 2$ % и до 22,6 % для $M = 20$ % (рис. 2, *g*).

Recall@100 в зависимости от процента известных белков в активном модуле при $\alpha = 0{,}5$ и числа вершин активного модуля $G = 50$. Прирост Recall@100 для предлагаемого метода по сравнению с методом на основе MCMC составил от 12 % для $M = 2$ % и до 43 % для $M = 20$ % (рис. 2, *h*).

Таким образом, показано улучшение качества предлагаемого метода по сравнению с методом на основе

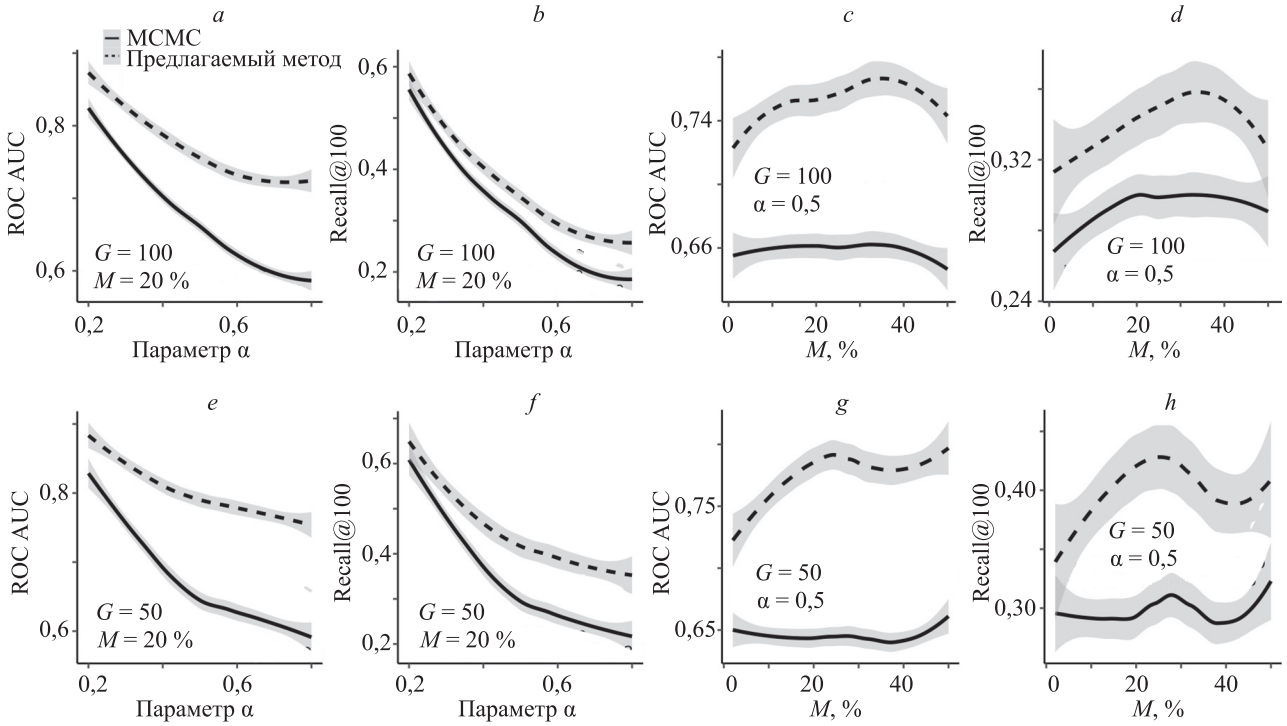

Рис. 2. Результаты симуляций для сети InWebIM. Сравнение метрик ROC AUC и Recall@100 в зависимости от параметра α бета-равномерного распределения и процента известных белков в активном модуле *M*.

Результаты симуляции для активных модулей с числом вершин *G*: 100 (*a*–*d*) и 50 (*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 percentage of known proteins in the active module *M*.

Simulation results for active modules with the number of vertices *G*: 100 (*a*–*d*) and 50 (*e*–*h*)

MCMC для активных модулей разных размеров, для различных параметров α бета-равномерного распределения и различного числа известных белков активного модуля.

Предлагаемый метод может быть применен к различным однородным биологическим сетям. В качестве примера выбрана независимая от сети InWebIM биологическая сеть функциональных ассоциаций генов GeneMANIA (13 386 вершин и 6 209 461 ребро) для вида живых организмов — Drosophila melanogaster [23]. Функциональные ассоциации генов в сети GeneMANIA экспериментально определялись на основе большого набора данных: белковые взаимодействия, метаболические пути, совместная локализация, схожесть белковых доменов. Вершинами в сети GeneMANIA выступают гены, а ребрами — наличие функциональной связи между парой генов.

Предлагаемый метод был применен для сети GeneMANIA с учетом того, что в вершинах находятся гены. Увеличение значений метрик ROC AUC и Recall@100 на сети GeneMANIA для предлагаемого метода по сравнению с методом на основе MCMC аналогичны изменениям этих метрик в сети InWebIM при сравнении рассматриваемых методов.

Таким образом, была показана устойчивость предлагаемого метода для различных биологических сетей (рис. 3).

Сравнение предлагаемого метода с методом на основе марковской цепи Монте-Карло при соблюдении всех предположений последнего

Математическая задача. Метод, основанный на MCMC, предполагает, что форма активного модуля может быть произвольной. Чтобы продемонстрировать независимость предлагаемого метода от предположений метода на основе MCMC, необходимо провести сравнение этих двух методов в условиях, когда предположения метода на основе MCMC выполняются.

Решение задачи. Используется предлагаемый метод, в котором алгоритм поиска в ширину при симулировании активных модулей заменяется на алгоритм случайного поиска, так как метод на основе MCMC предполагает произвольную форму активных модулей. Активные модули выбираются равновероятно среди всех возможных связных подграфов заданного размера с использованием mcmcRanking (v0.1.0) [20] при условии, что все вершины имеют одинаковый вес. Модули выбираются после 1000 итераций MCMC, чтобы гарантировать их независимость от подграфов, использованных для инициализации MCMC. Далее предлагаемый метод применяется к сети InWebIM.

Результаты. Результаты показали, что как ROC AUC, так и Recall@100 незначимо отличались между предлагаемым методом и методом на основе MCMC

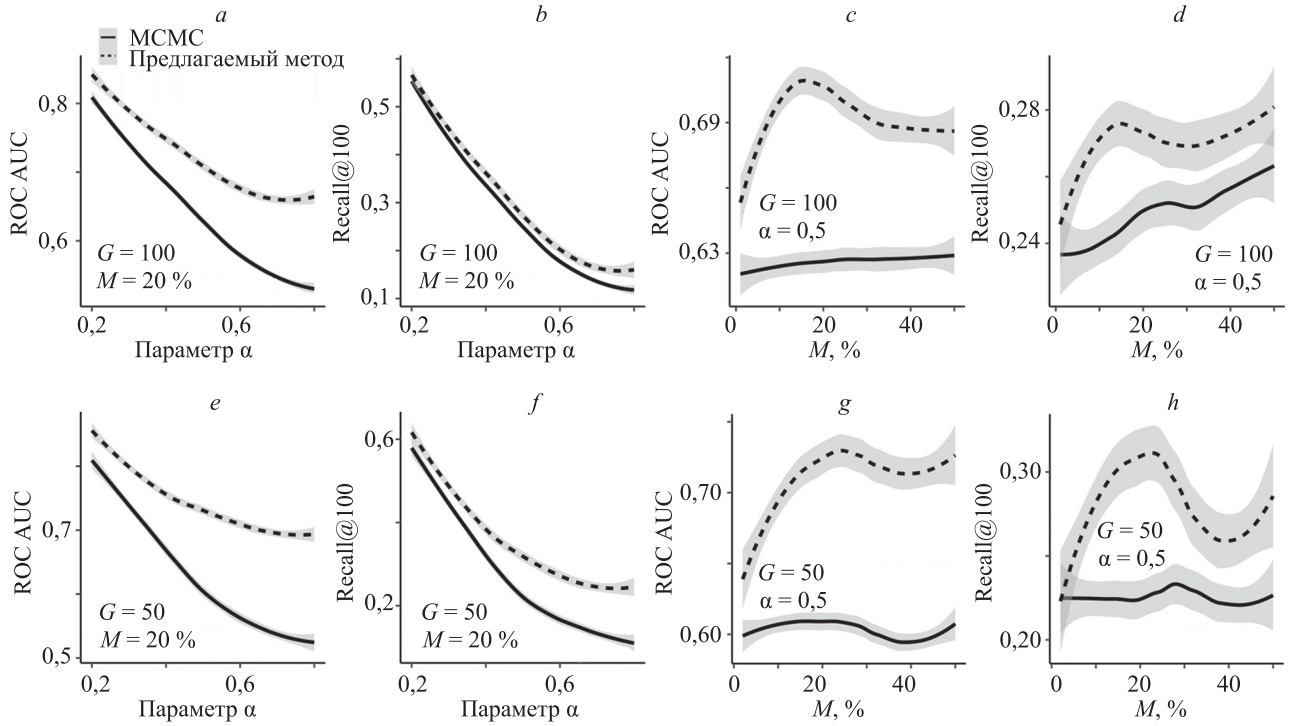

Рис. 3. Результаты симуляций для GeneMANIA (Drosophila melanogaster). Сравнение метрик ROC AUC и Recall@100 в зависимости от параметра α бета-равномерного распределения и процента известных генов в активном модуле *M*.

Результаты симуляции для активных модулей с числом вершин *G*: 100 (*a*–*d*) и 50 (*e*–*h*)

Fig. 3. Simulation results for GeneMANIA (Drosophila melanogaster). Comparison of ROC AUC and Recall@100 metrics vs. the parameter α of the beta-uniform distribution, the percentage of known genes in the active module, and the size of the active module *M*.

Simulation results for active modules with the number of vertices *G*: 100 (*a*–*d*) and 50 (*e*–*h*)

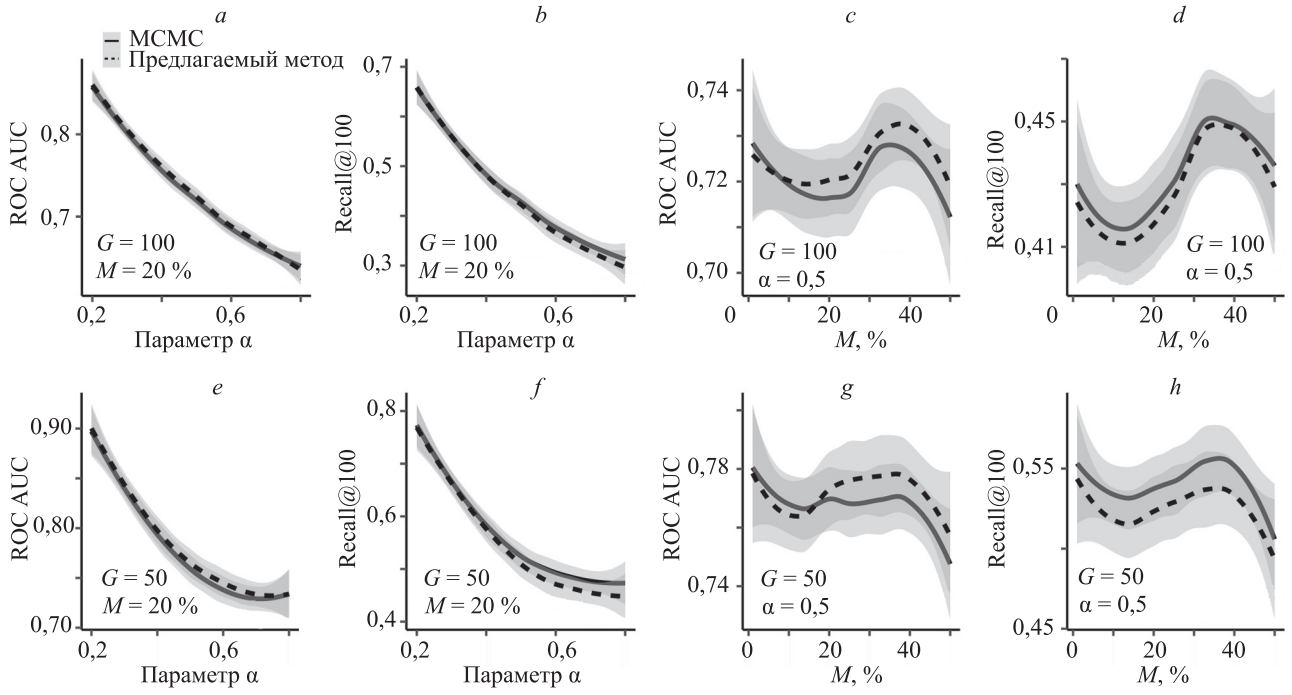

Рис. 4. Результаты для сети InWebIM с использованием случайного поиска при симуляции активных модулей. Сравнение метрик ROC AUC и Recall@100 в зависимости от параметра α бета-равномерного распределения и процента известных генов в активном модуле *M*.

Результаты симуляции для активных модулей с числом вершин *G*: 100 (*a*–*d*) и 50 (*e*–*h*)

Fig. 4. Results for InWebIM using random search for active module simulations. Comparison of ROC AUC and Recall@100 metrics vs. the parameter α of the beta-uniform distribution, the percentage of known genes in the active module, and the size of the active module *M*.

Simulation results for active modules with the number of vertices *G*: 100 (*a*–*d*) and 50 (*e*–*h*)

(рис. 4). Таким образом, предлагаемый метод работает стабильно не хуже метода на основе MCMC, даже в случае, когда все предположения метода на основе MCMC верны. В случае, когда предположения метода на основе MCMC не соблюдаются, например, активные модули генерируются поиском в ширину, предлагаемый метод показывает значимое повышение точности нахождения активных модулей по сравнению с методом на основе MCMC.

Заключение

В работе предложен метод повышения точности определения активных модулей в биологических графах за счет последовательного использования метода на основе марковской цепи Монте-Карло (MCMC) и метода на основе модели машинного обучения — градиентного бустинга. Для исследования устойчивости предложенного метода, проведено его сравнение на сотнях активных модулей с использованием различных параметров. Применимость предложенного метода к различным биологическим сетям, была продемонстрирована на примере сети белок-белковых взаимодействий InWebIM и сети функциональных ассоциаций генов для вида живых организмов — Drosophila melanogaster. В случае, когда активные модули формируются случайным поиском, предлагаемый метод работает не хуже, чем метод на основе MCMC. В случае, когда активные модули формируются поиском в ширину, предлагаемый метод работает значительно лучше, чем метод на основе MCMC. Так как заранее неизвестно, каким образом активные модули генерируются в природе, предлагаемый метод является более устойчивым и практичным в применении.

Показано, что предлагаемый подход может повышать точность определения активного модуля даже в случаях, когда заранее известно лишь небольшое число вершин активного модуля.

Литература

1. Huber W., Carey V.J., Long L., Falcon S., Gentleman R. Graphs in molecular biology // BMC Bioinformatics. 2007. V. 8. Suppl. 6. P. S8. https://doi.org/10.1186/1471-2105-8-S6-S8
2. Szczepanski A.P., Wang L. Emerging multifaceted roles of BAP1 complexes in biological processes // Cell Death Discovery. 2021. V. 7. N 1. P. 20. https://doi.org/10.1038/s41420-021-00406-2
3. Carbone M., Yang H., Pass H.I., Krausz T., Testa J.R., Gaudino G. BAP1 and cancer // Nature Reviews Cancer. 2013. V. 13. N 3. P. 153–159. https://doi.org/10.1038/nrc3459
4. Lin J.S., Lai E.M. Protein-protein interactions: Co-Immunoprecipitation // Methods in Molecular Biology. 2017. V. 1615. P. 211–219. https://doi.org/10.1007/978-1-4939-7033-9_17
5. Tamara S., den Boer M.A., Heck A.J.R. High-resolution native mass spectrometry // Chemical Reviews. 2022. V. 122. N 8. P. 7269–7326. https://doi.org/10.1021/acs.chemrev.1c00212
6. Okpara M.O., Hermann C., van der Watt P.J., Garnett S., Blackburn J.M., Leaner V.D. A mass spectrometry-based approach for the identification of Kpnβ1 binding partners in cancer cells // Scientific Reports. 2022. V. 12. N 1. P. 20171. https://doi.org/10.1038/s41598-022-24194-6
7. 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
8. Zhu Q.M., Hsu Y.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
9. Nehme R., Pietiläinen O., Artomov M., Tegtmeyer M., Valakh V., Lehtonen L., Bell C., Singh T., Trehan A., Sherwood J., Manning D., Peirent E., Malik R., Guss E.J., Hawes D., Beccard A., Bara A.M., Hazelbaker D.Z., Zuccaro E., Genovese G., Loboda A.A., Neumann A., Lilliehook C., Kuismin O., Hamalainen E., Kurki M., Hultman C.M., Kähler A.K., Paulo J.A., Ganna A., Madison J., Cohen B., McPhie D., Adolfsson R., Perlis R., Dolmetsch R., Farhi S., McCarroll S., Hyman S., Neale B., Barrett L.E., Harper W., Palotie A., Daly M., Eggan K. 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
10. Nguyen H., Shrestha S., Tran D., Shafi A., Draghici S., Nguyen T. A Comprehensive survey of tools and software for active subnetwork

References

1. Huber W., Carey V.J., Long L., Falcon S., Gentleman R. Graphs in molecular biology. *BMC Bioinformatics*, 2007, vol. 8, suppl. 6, pp. S8. https://doi.org/10.1186/1471-2105-8-S6-S8
2. Szczepanski A.P., Wang L. Emerging multifaceted roles of BAP1 complexes in biological processes. *Cell Death Discovery*, 2021, vol. 7, no. 1, pp. 20. https://doi.org/10.1038/s41420-021-00406-2
3. Carbone M., Yang H., Pass H.I., Krausz T., Testa J.R., Gaudino G. BAP1 and cancer. *Nature Reviews Cancer*, 2013, vol. 13, no. 3, pp. 153–159. https://doi.org/10.1038/nrc3459
4. Lin J.S., Lai E.M. Protein-protein interactions: Co-Immunoprecipitation. *Methods in Molecular Biology*, 2017, vol. 1615, pp. 211–219. https://doi.org/10.1007/978-1-4939-7033-9_17
5. Tamara S., den Boer M.A., Heck A.J.R. High-resolution native mass spectrometry. *Chemical Reviews*, 2022, vol. 122, no. 8, pp. 7269–7326. https://doi.org/10.1021/acs.chemrev.1c00212
6. Okpara M.O., Hermann C., van der Watt P.J., Garnett S., Blackburn J.M., Leaner V.D. A mass spectrometry-based approach for the identification of Kpnβ1 binding partners in cancer cells. *Scientific Reports*, 2022, vol. 12, no. 1, pp. 20171. https://doi.org/10.1038/s41598-022-24194-6
7. 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
8. Zhu Q.M., Hsu Y.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
9. Nehme R., Pietiläinen O., Artomov M., Tegtmeyer M., Valakh V., Lehtonen L., Bell C., Singh T., Trehan A., Sherwood J., Manning D., Peirent E., Malik R., Guss E.J., Hawes D., Beccard A., Bara A.M., Hazelbaker D.Z., Zuccaro E., Genovese G., Loboda A.A., Neumann A., Lilliehook C., Kuismin O., Hamalainen E., Kurki M., Hultman C.M., Kähler A.K., Paulo J.A., Ganna A., Madison J., Cohen B., McPhie D., Adolfsson R., Perlis R., Dolmetsch R., Farhi S., McCarroll S., Hyman S., Neale B., Barrett L.E., Harper W., Palotie A., Daly M., Eggan K. 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
10. Nguyen H., Shrestha S., Tran D., Shafi A., Draghici S., Nguyen T. A Comprehensive survey of tools and software for active subnetwork

identification // Frontiers in Genetics. 2019. V. 10. P. 155. https://doi.org/10.3389/fgene.2019.00155
11. Mitra K., Carvunis A.R., Ramesh S.K., Ideker T. Integrative approaches for finding modular structure in biological networks // Nature Reviews Genetics. 2013. V. 14. N 10. P. 719–732. https://doi.org/10.1038/nrg3552
12. Strauss B.S. Biochemical genetics and molecular biology: The contributions of George Beadle and Edward Tatum // Genetics. 2016. V. 203. N 1. P. 13–20. https://doi.org/10.1534/genetics.116.188995
13. Montecino-Rodriguez E., Casero D., Fice M., Le J., Dorshkind K. Differential expression of PU.1 and key T lineage transcription factors distinguishes fetal and adult T cell development // Journal of Immunology. 2018. V. 200. N 6. P. 2046–2056. https://doi.org/10.4049/jimmunol.1701336
14. Suzuki K., Hatzikotoulas K., Southam L., Taylor H.J., Yin X., Lorenz K.M. et al. Genetic drivers of heterogeneity in type 2 diabetes pathophysiology // Nature. 2024. V. 627. P. 347–357. https://doi.org/10.1038/s41586-024-07019-6
15. 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
16. Barton S.J., Crozier S.R., Lillycrop K.A., Godfrey K.M., Inskip H.M. Correction of unexpected distributions of P values from analysis of whole genome arrays by rectifying violation of statistical assumptions // BMC Genomics. 2013. N 14. P. 161. https://doi.org/10.1186/1471-2164-14-161
17. 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
18. 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
19. 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
20. Chen T., Guestrin C. XGBoost: A scalable tree boosting system // Proc. of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016. P. 785–794. https://doi.org/10.1145/2939672.2939785
21. Warde-Farley D., Donaldson S.L., Comes O., Zuberi K., Badrawi R., Chao P., Franz M., Grouios C., Kazi F., Lopes C.T., Maitland A., Mostafavi S., Montojo J., Shao Q., Wright G., Bader G.D., Morris Q. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function // Nucleic Acids Research. 2010. V. 38. Suppl. 2. P. W214–W220. https://doi.org/10.1093/nar/gkq537

identification. *Frontiers in Genetics*, 2019, vol. 10, pp. 155. https://doi.org/10.3389/fgene.2019.00155
11. Mitra K., Carvunis A.R., Ramesh S.K., Ideker T. Integrative approaches for finding modular structure in biological networks. *Nature Reviews Genetics*, 2013, vol. 14, no. 10, pp. 719–732. https://doi.org/10.1038/nrg3552
12. Strauss B.S. Biochemical genetics and molecular biology: The contributions of George Beadle and Edward Tatum. *Genetics*, 2016, vol. 203, no. 1, pp. 13–20. https://doi.org/10.1534/genetics.116.188995
13. Montecino-Rodriguez E., Casero D., Fice M., Le J., Dorshkind K. Differential expression of PU.1 and key T lineage transcription factors distinguishes fetal and adult T cell development. *Journal of Immunology*, 2018, vol. 200, no. 6, pp. 2046–2056. https://doi.org/10.4049/jimmunol.1701336
14. Suzuki K., Hatzikotoulas K., Southam L., Taylor H.J., Yin X., Lorenz K.M. et al. Genetic drivers of heterogeneity in type 2 diabetes pathophysiology. *Nature*, 2024, vol. 627, pp. 347–357. https://doi.org/10.1038/s41586-024-07019-6
15. 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
16. Barton S.J., Crozier S.R., Lillycrop K.A., Godfrey K.M., Inskip H.M. Correction of unexpected distributions of P values from analysis of whole genome arrays by rectifying violation of statistical assumptions. *BMC Genomics*, 2013, no. 14. pp. 161. https://doi.org/10.1186/1471-2164-14-161
17. 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
18. 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
19. 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
20. Chen T., Guestrin C. XGBoost: A scalable tree boosting system. *Proc. of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794. https://doi.org/10.1145/2939672.2939785
21. Warde-Farley D., Donaldson S.L., Comes O., Zuberi K., Badrawi R., Chao P., Franz M., Grouios C., Kazi F., Lopes C.T., Maitland A., Mostafavi S., Montojo J., Shao Q., Wright G., Bader G.D., Morris Q. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Research*, 2010, vol. 38, suppl. 2, pp. W214–W220. https://doi.org/10.1093/nar/gkq537

Авторы

Усольцев Дмитрий Андреевич — аспирант, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация; старший научный сотрудник, Институт геномной медицины, Детская больница Нейшенвайд, Колумбус, 43205, США, 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, 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

Authors

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

Ivan I. Molotkov — Senior Researcher, Institute for Genomic Medicine, Nationwide Chldren'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 Chldren's Hospital, Columbus, 43205, USA; Professor, 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

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

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

Статья поступила в редакцию 10.09.2024
Одобрена после рецензирования 02.10.2024
Принята к печати 15.11.2024

Received 10.09.2024
Approved after reviewing 02.10.2024
Accepted 15.11.2024