УНИВЕРСИТЕТ ИТМО
НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
ноябрь–декабрь 2020 Том 20 № 6 ISSN 2226-1494 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
November–December 2020 Vol. 20 No 6 ISSN 2226-1494 http://ntv.ifmo.ru/en/
НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК УНИВЕРСИТЕТ ИТМО ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

УДК 519.1 doi: 10.17586/2226-1494-2020-20-6-807-814

МОДИФИКАЦИЯ МЕТОДА СОВМЕСТНОЙ КЛАСТЕРИЗАЦИИ В ГРАФОВОМ И КОРРЕЛЯЦИОННОМ ПРОСТРАНСТВАХ

А.Н. Гайнуллина, М. Артемов, А.А. Сергушичев

Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация
Адрес для переписки: anastasiia.gainullina@gmail.com

Информация о статье
Поступила в редакцию 13.09.20, принята к печати 31.10.20
Язык статьи — русский

Ссылка для цитирования: Гайнуллина А.Н., Артемов М., Сергушичев А.А. Модификация метода совместной кластеризации в графовом и корреляционном пространствах // Научно-технический вестник информационных технологий, механики и оптики. 2020. Т. 20. № 6. С. 807–814. doi: 10.17586/2226-1494-2020-20-6-807-814

Аннотация

Предмет исследования. Метод совместной кластеризации в графовом и корреляционном пространствах предназначен для идентификации активных модулей в метаболических графах на основании транскриптомных данных, представленных большим числом образцов. Получаемые с помощью данного метода активные модули описывают динамическую метаболическую регуляцию во всех образцах анализируемого набора данных. В работе изложены способы модификации предложенного метода для применения на реальных данных. **Метод.** Для изучения устойчивости получаемых результатов модифицированный метод был многократно запущен на реальных данных с небольшими вариациями исходных параметров. Для анализа результатов сформулировано несколько метрик, отражающих степень похожести и представленности полученных при разных запусках модулей. **Основные результаты.** Результаты анализа в целом являются достаточно устойчивыми: для большинства модулей их профили хорошо находятся в шумных данных, а также сохраняется большинство генов этих модулей. **Практическая значимость.** Результаты приведенных исследований показали, что используемые модификации метода позволяют успешно анализировать реальные данные путем получения активных модулей, обладающих устойчивостью и простотой в интерпретации.

Ключевые слова
кластеризация, корреляция, графы, метаболические сети, экспрессия генов, транскриптомные данные

Благодарности
Работа выполнена при поддержке Правительства Российской Федерации, субсидия 08-08.

doi: 10.17586/2226-1494-2020-20-6-807-814

METHOD OF JOINT CLUSTERING IN NETWORK AND CORRELATION SPACES

A.N. Gainullina, M. Artyomov, A.A. Sergushichev

ITMO University, Saint Petersburg, 197101, Russian Federation
Corresponding author: anastasiia.gainullina@gmail.com

Article info
Received 13.09.20, accepted 31.10.20
Article in Russian

For citation: Gainullina A.N., Artyomov M., Sergushichev A.A. Method of joint clustering in network and correlation spaces. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2020, vol. 20, no. 6, pp. 807–814 (in Russian). doi: 10.17586/2226-1494-2020-20-6-807-814

Abstract

Subject of Research. The joint clustering method in network and correlation context is designed to identify active modules in metabolic graphs based on transcriptomic data represented by a large number of samples. The active modules obtained by this method describe the dynamic metabolic regulation in all samples of the analyzed dataset. The paper presents modifications of the proposed method for application on real data. **Method.** For results stability study the modified method was repeatedly run on real data with small variations of the initial parameters. For result analysis, several metrics were formulated that display modules similarity and representation under various start-up conditions. **Main Results.** The analysis results are sufficiently robust. For the most modules, their profiles are detected well in the noisy data, and the most genes are also preserved. **Practical Relevance.** The results of the presented study have shown that the modified method analyzes successfully real data by producing active modules that are stable and easy in interpretation.

Keywords
clustering, correlation, graphs, metabolic networks, gene expression, transcriptomic data

Acknowledgements
This work was supported by the Government of the Russian Federation, Investigation Research Grant 08-08.

Введение

Изучение обмена веществ, или метаболизма является одним из важных и перспективных направлений исследований в биологии [1–3]. Метаболизм представляет собой набор биохимических реакций, протекающих в организме. Одним из способов систематического анализа метаболизма является изучение данных, получаемых в результате так называемого *транскриптомного профилирования* (*транскриптомных данных*) [4, 5]. С физической точки зрения этот анализ позволяет оценить то, насколько активны те или иные биохимические реакции через уровни накопления (*экспрессии*) соответствующих им генов. С технической точки зрения результатом данного анализа является *таблица экспрессии генов*, строки которой соответствуют собственно генам, а столбцы — изучаемым *образцам* (например, клеткам из разных тканей или в разных биологических состояниях). В ячейках данной таблицы находятся числа, характеризующие уровни экспрессии конкретных генов в конкретных образцах.

Изучение таблицы экспрессии генов позволяет судить о том, как различаются активности биохимических реакций в разных образцах. Ранее было показано [6–8], что анализ транскриптомных данных с использованием графов улучшает интерпретируемость результатов из-за наличия естественной графовой структуры метаболических реакций. В частности, все метаболические реакции группируются в метаболическую сеть, которую можно представить в виде графа, вершины которого соответствуют генам, ответственным за превращение одних веществ в другие (будем называть этот граф *метаболическим графом*). Однако существующие на данный момент методы [9] используются лишь для парных сравнений образцов (т. е. работают с таблицами экспрессии, состоящими лишь из двух колонок) и напрямую не подходят для анализа экспериментов, состоящих из многих образцов. Подходом к расширению графовых методов на работу с данными, состоящими из многих образцов, может являться интеграция этих методов с методами кластеризации.

В настоящей работе рассматривается использование предложенного ранее метода совместной кластеризации в графовом и корреляционном пространствах для анализа активности метаболических процессов в образцах транскриптомных данных [10]. Целью этого метода является идентификация связных участков (подграфов) метаболического графа, активность которых регулируется совместно от образца к образцу. Такие подграфы будем называть *активными модулями* [11].

Постановка проблемы

Первым этапом метода совместной кластеризации в графовом и корреляционном пространствах является предварительная кластеризация значений экспрессии n генов (строк таблицы экспрессии генов) по корреляции с помощью метода *k-medoids* при заданном значении числа начальных кластеров k. Затем для каждого полученного кластера вычисляется так называемый *профиль* — усредненное значение экспрессии всех входящих в кластер генов.

Далее для каждого гена g_i ($i \in [1\ldots n]$) таблицы вычисляется корреляционное расстояние Пирсона до каждого из k таких начальных профилей c_j ($j \in [1\ldots k]$):

$$d(g_i, c_j) = 1 - \mathrm{corr}(g_i, c_j).$$

Полученный массив значений формирует матрицу расстояний, где каждая строка соответствует какому-либо гену, а каждый столбец — конкретному профилю начальных кластеров.

Данные значения используются для вычисления наборов генных весов для каждого из кластеров.

Для этого, во-первых, введем фиктивный нулевой профиль c_0, расстояние до которого по определению будет всегда равно некоторой константе *base*:

$$d(g_i, c_0) \equiv base.$$

Во-вторых, по формуле

$$d'(g_i, c_j) = \min_{l \neq j, l \in [0\ldots k]} d(g_i, c_l)$$

также для всех пар генов g_i ($i \in [1\ldots n]$) и профилей c_j ($j \in [1\ldots k]$) найдем расстояние до ближайшего профиля (без учета расстояния до профиля, для которого рассчитывается расстояние d'). При этом поиске расстояния до ближайшего профиля будем также учитывать расстояние до фиктивного нулевого профиля c_0, равное значению *base*.

В итоге, по формуле

$$\mathrm{weight}(g_i, c_j) = -\log \frac{d(g_i, c_j)}{d'(g_i, c_j)}$$

получим k наборов значений генных весов, на основании которых происходит поиск k подграфов максимального веса в метаболическом графе — собственно метаболических модулей.

Нетрудно видеть, что вес может быть положительным, только если корреляция между экспрессией g_i гена и c_j профилем больше $1 - base$.

Итак, для каждого найденного модуля его гены, имеющие положительный вес, становятся новыми кластерами. Обновленный список генов кластера влечет за собой вычисление нового профиля, что, в свою очередь, приводит к обновлению генных весов. Эта процедура продолжается до тех пор, пока изменение генного состава кластеров не становится пренебрежимо мало.

Важными параметрами метода являются число начальных кластеров k, а также параметр *base*, регулирующий число генов с положительным весом в кластере, задавая минимальный допустимый уровень корреляции экспрессии гена с профилем. В настоящей статье рассматривается проблема применения метода совместной кластеризации в графовом и корреляционном пространствах к реальным транскриптомным данным, а именно, исследуется автоматический выбор параметров метода и стабильность получаемых результатов.

Описание использованных данных

В настоящей работе использовалось три реальных набора данных для изучения работы метода совместной

кластеризации в графовом и корреляционном пространствах:

1) IG.OS (ImmGen Open Source [12]) — данные транскриптомного профилирования мышиных мононуклеарных фагоцитов из разных тканей;
2) IG.P1 (ImmGen Phase1 [13]) — данные транскриптомного профилирования мышиных иммунных клеток из разных тканей, из которых также была взята подгруппа мононуклеарных фагоцитов;
3) EVI (Endovascular Intervention [14]) — данные транскриптомного профилирования образцов крысиных сосудов, подвергшихся экспериментальному повреждению в рамках изучения динамики заживления сосудов (данные представляют собой разные временны́е точки одного и того же биологического процесса).

Выбор параметра, регулирующего число генов с положительным весом в кластере

Как было показано в предыдущей работе авторов [10], для больших значений *base* метод выдает результаты, характеризующиеся большей полнотой, но меньшей точностью. В связи с этим однозначно выбрать какое-то оптимальное значение этого параметра затруднительно. Однако можно предложить способ автоматического динамического подбора его значения на основе входных данных для достижения удобных для интерпретации результатов.

Предлагаемый способ подбора параметра *base* основан на предположении, что для интерпретируемости модули не должны иметь размер больше *maxSize* = = 50 генов. Для выполнения этого условия размеры полученных на *i*-ой итерации подграфов сравниваются с величиной *maxSize*. В случае, если какой-то подграф имеет более 50 уникальных генов, на новой итерации величина *base* уменьшается на некоторую заданную величину *baseDecrement* = 0,05.

Стоит отметить, что изначально величина *base* задается равной 1, что соответствует нулевой корреляции между профилем кластера и геном, а ее уменьшение означает все более и более высокие требования к их скоррелированности.

Псевдокод модифицированного метода приведен на рис. 1.

Algorithm: Network clustering

Input: Graph $G = (V, E)$ of order $n = |V|$, matrix X of size $n \times m$, initial module profiles approximation $P^{(1)}$ of size $k^{(1)} \times m$, value of maximum module size limit *maxSize*, initial value of *base*, value of *baseDecrement*.

Result: Final approximation of profiles P^* of size $k^* \times m$ and a set of connected subgraphs A_i^* for $i \in 1, \ldots, k^*$ as a final approximation of active modules.

for $i \in \{1, 2, \ldots\}$ **do**
  $k^{(i)} \leftarrow$ number of rows in $P^{(i)}$;
  $d_{x,y} \leftarrow 1 - \mathrm{corr}(X_x, P_y^{(i)})$ for $x \in \{1, \ldots, n\}, y \in \{1, \ldots, k^{(i)}\}$;
  $d_{x,0} \leftarrow base$ for $x \in \{1, \ldots, n\}$;
  $d'_{x,y} \leftarrow \min_{z \in \{0, \ldots, k^{(i)}\}, z \neq y} d_{x,z}$ for $x \in \{1, \ldots, n\}, y \in \{1, \ldots, k^{(i)}\}$;
  for $j \in \{1, \ldots, k^{(i)}\}$ **do**
    $w_x \leftarrow -\log \frac{d_{x,y}}{d'_{x,y}}$ for $x \in \{1, \ldots, n\}$;
    $A_j^{(i)} \leftarrow$ connected subgraph of G with maximum sum of vertex weights w ;
    $P^{(i+1)} \leftarrow$ coordinate-wise average of X_x, for $x \in V(A_j^{(i)})$ if $w_x > 0$;
  end
  if $\max_{j \in \{1, \ldots, k^{(i)}\}} |E(A_j^{(i)})| > maxSize$ **then**
    $base \leftarrow base - baseDecrement$;
    continue
  end
  if $P^{(i+1)}$ *substantially differs from* $P^{(j)}$ *for* $j \leq i$ **then**
    continue
  end
  if *there are very small modules in* $A^{(i)}$ **then**
    remove one row from $P^{(i+1)}$ that corresponds to the smallest module ;
    continue
  end
  break
end

Рис. 1. Модифицированный метод совместной кластеризации в графовом и корреляционном пространствах с динамическим вычислением параметра *base*

Выбор параметра, определяющего число начальных кластеров

Другим параметром метода является число исходных кластеров *k*.

Согласно результатам экспериментов на симулированных данных, показанным в [10], где истинное число модулей было заранее известно, достижение хороших стартовых приближений и итоговых показателей возможно при значениях *k*, в несколько раз больших истинного числа модулей, причем результаты улучшаются с увеличением *k*. В среднем для достижения оптимальных результатов требуется число начальных кластеров примерно в три-четыре раза большее, чем истинное число модулей. При этом число итераций и время, затраченное методом на поиск активных модулей, растет нелинейно в зависимости от изначального числа кластеров, полученных на первом этапе метода, посвященному предварительной кластеризации.

Таким образом, необходимо найти компромисс между достаточно высоким значением *k* при неизвестном числе модулей, не сделав при этом время работы метода чрезмерно большим.

Для нахождения оптимального числа кластеров в каком-либо наборе данных рекомендуется строить график зависимости между числом кластеров и общей внутрикластерной вариацией (метрика WSS, within-cluster sum of squares) для каждого конкретного значения *k*. Согласно эвристическому методу *«локтя»*, с помощью такого графика можно определить оптимальное число кластеров по точке, в которой внутрикластерная вариация резко перестает уменьшаться [15]. Однако стоит отметить, что оптимальное число кластеров во входных транскриптомных данных не тождественно истинному числу активных модулей. В связи с этим метод «локтя» можно использовать лишь как способ приблизительной оценки параметра *k*.

На рис. 2 представлены графики, построенные по методу «локтя», для трех рассматриваемых наборов транскриптомных данных. По представленным кривым видно, что для всех трех наборов данных значение общей внутрикластерной вариации перестает резко уменьшаться при значениях *k* примерно в диапазоне 8–16. Таким образом, можно полагать, что значение 48 (в три раза большее, чем 16) можно использовать в качестве числа начальных кластеров для этих наборов данных.

Значение $k = 48$ также было выбрано как значение по умолчанию. Во-первых, все три транскриптомных набора данных показали однотипное поведение на приведенном графике, и можно ожидать похожих графиков и от других наборов данных. С другой стороны, при числе финальных модулей большем 10–15 их становится сложно интерпретировать. Наконец, наличие значения по умолчанию позволяет не строить график внутрикластерной вариации при каждом запуске, что значительно сокращает время работы метода.

Анализ устойчивости результатов метода с помощью варьирования параметров входных данных при запуске на реальных данных

Оценка точности предложенного метода на реальных данных затруднена, так как в данном случае неизвестны ни сами модули, ни их число. Однако для оценки качества метода можно провести анализ его устойчивости при вариациях в исходных данных.

Для анализа устойчивости каждый из трех рассматриваемых наборов данных был проанализирован с помощью предложенного метода с динамическим выбором параметра *base* и при значении $k = 48$. Этот анализ и его результаты будем называть *референсными*.

Далее для каждого из набора данных запускалось два типа экспериментов. Во-первых, производился анализ при небольшом варьировании параметра k с присвоением ему одного из значений из множества {46, 47, 49, 50}. Во-вторых, производился анализ при $k = 48$, но с добавлением небольшого шума из нормального распределения со значением среднеквадратичного отклонения, равным 0,05, ко всем значениям экспрессии генов; этот эксперимент проводился пять раз для разных исходных значений состояния генератора случайных чисел.

На рис. 3 для этих запусков приведены финальные значения параметра *base*, а также число найденных методом модулей. Несмотря на наличие некоторой вариации в результатах, метод показывает достаточную устойчивость относительно значения, полученного при референсном анализе.

Затем модули, полученные при референсном анализе, сравнивались с модулями, полученными в экспериментах с небольшими отличиями входных данных, согласно следующей процедуре.

Сначала для профиля каждого референсного модуля вычислялась его корреляция с профилями модулей, полученных в каждом из девяти экспериментов с отличиями во входных данных (четыре эксперимента с вариацией k и пять экспериментов с добавлением шума).

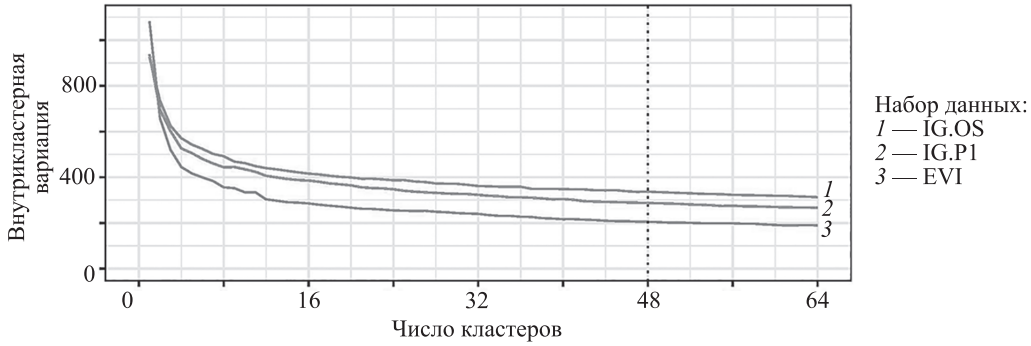

Рис. 2. График, построенный по методу «локтя», для реальных транскриптомных данных; пунктирной линией показано выбранное значение по умолчанию для параметра k

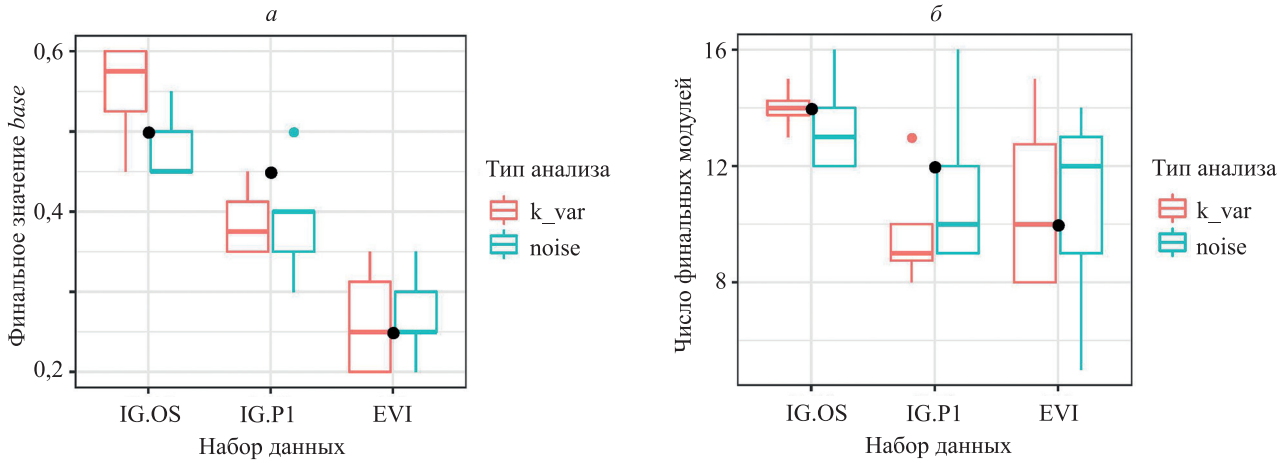

Рис. 3. Финальные значения *base* (*а*) и число найденных модулей (*б*), получившиеся при разных экспериментах: k_var — вариация значения параметра k, noise — добавление небольшого шума из нормального распределения к данным экспрессии; черной точкой обозначены значения для референсного анализа каждого из набора данных

Затем для каждого референсного модуля выбиралось максимальное значение среди корреляций с профилями финальных модулей каждого из этих девяти экспериментов (всего девять значений для каждого модуля). Это значение показывает, насколько хорошо профиль референсного модуля коррелирует с наиболее похожим профилем в каждом из экспериментов с отличиями во входных данных (*maxCor*).

Кроме этого, для генов с положительным весом каждого из референсных модулей вычислялись еще две метрики: *retainedGenes* — процент генов, сохранившихся во всех модулях для каждого из девяти экспериментов с отличиями во входных данных, и *maxInter* — максимальный размер пересечения с генами положительного веса модулей в каждом из девяти вышеобозначенных экспериментов, деленное на число генов референсного модуля с положительным весом (также по девять значений каждой из метрик для каждого референсного модуля).

Как видно из определений, чем больше значение каждой из сформулированных метрик, тем более устойчивы полученные при референсном анализе модули. При этом ожидается, что модули большего размера будут более устойчивы, чем модули небольшого размера (содержащие семь и менее генов) ввиду неизбежной зашумленности реальных транскриптомных данных.

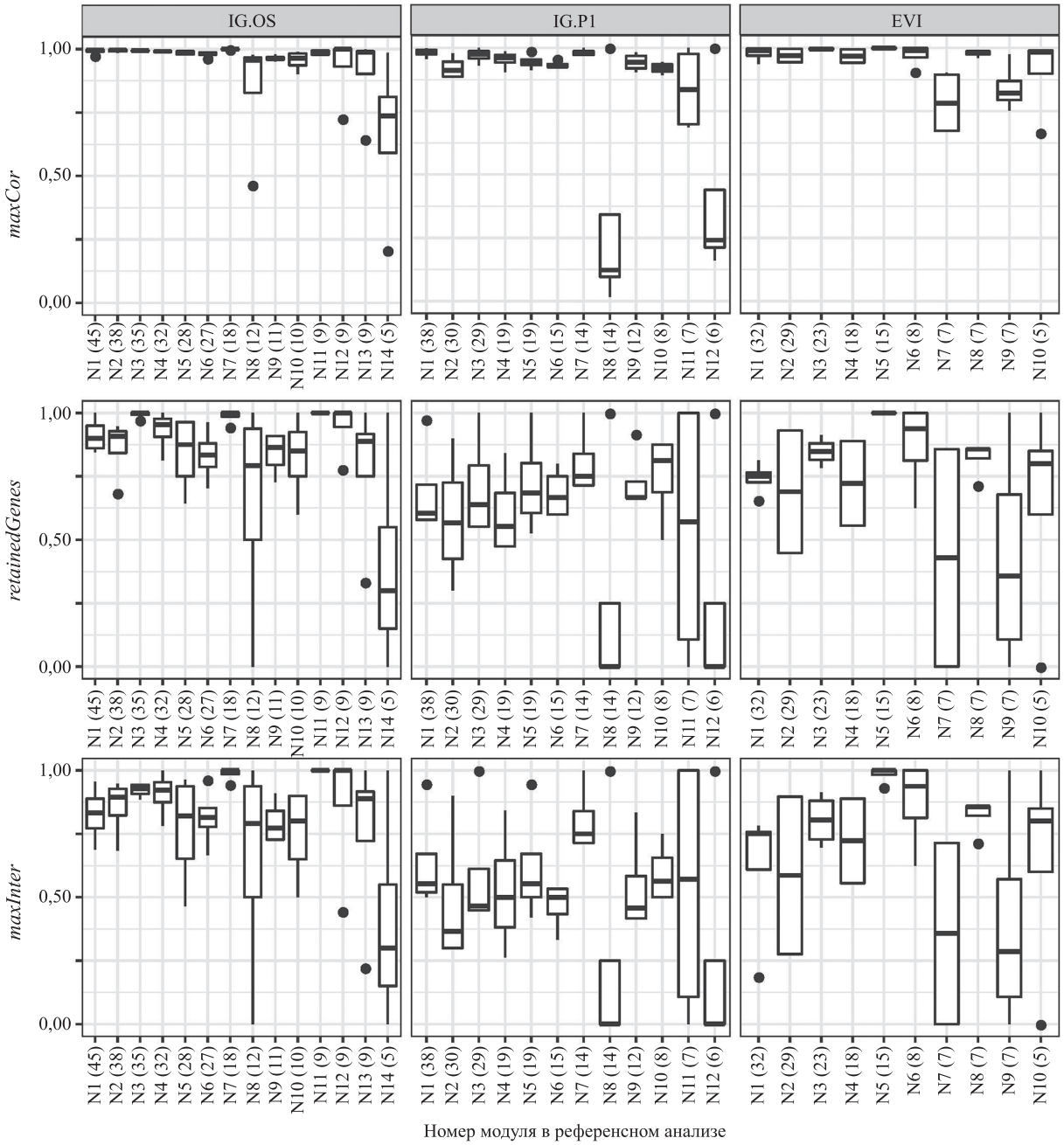

Рис. 4. Метрики стабильности модулей, полученные при сравнении результатов референсного анализа с результатами экспериментов с небольшими отклонениями параметра *k* (каждый бокс-плот сформирован четырьмя точками для *k* из множества {46, 47, 49, 50}, референсный анализ соответствует значению *k* = 48)

На рис. 4 и 5 показаны результаты двух вышеописанных типов экспериментов, представленные с помощью сформулированных метрик. Модули, номера которых отложены по оси абсцисс, являются результатами, полученными на следующих наборах данных: IG.OS — ImmGen Open Source, IG.P1 — ImmGen Phase1, EVI — Endovascular Intervention; рядом с номером каждого модуля в скобках указано число его генов с положительным весом.

На рис. 4 приведены результаты сравнения результатов референсного анализа с результатами экспериментов с небольшими отклонениями параметра *k*. По результатам, представленным на графике, видно, что значение *maxCor* высоко для большинства модулей, что говорит об устойчивости выводящихся методом профилей. Только два модуля из набора данных IG.P1 показали среднюю корреляцию меньше, чем 0,7. Показатели метрик *retainedGenes* и *maxInter* схожи друг с другом по поведению. Их значения довольно сильно различаются между наборами данных, что свидетельствует о разной стабильности модулей, полученных на разных таблицах генной экспрессии. В среднем, для набора данных IG.OS сохраняется более 75 % генов модулей, для наборов данных IG.P1 и EVI — более 50 %. Стоит

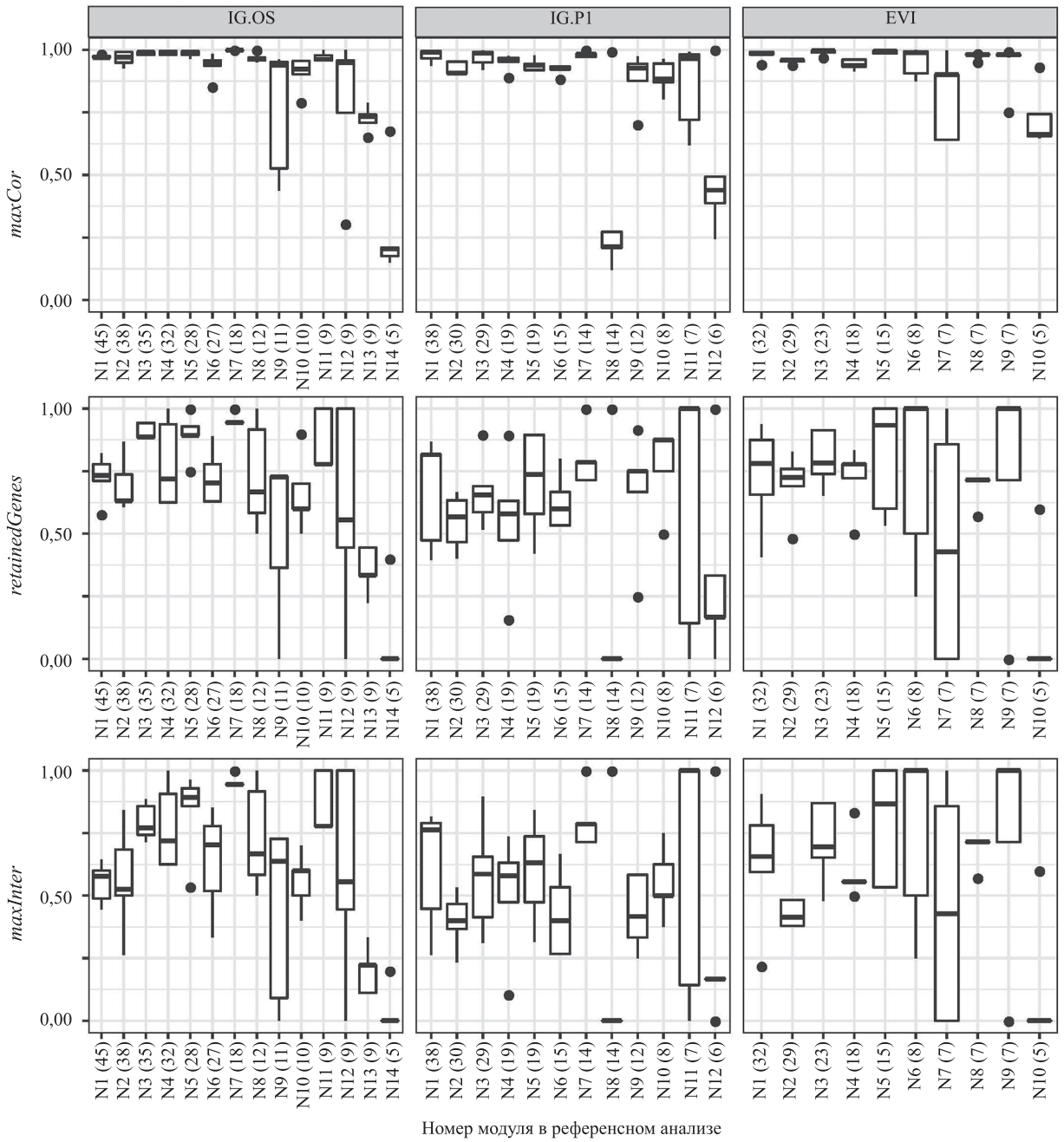

Рис. 5. Метрики стабильности модулей, полученные при сравнении результатов референсного анализа с результатами экспериментов с добавлением небольшого шума из нормального распределения со значением среднеквадратичного отклонения, равным 0,05, ко всем значениям экспрессии генов (каждый бокс-плот сформирован пятью точками для разных исходных значений состояния генератора случайных чисел, все эксперименты запущены при $k = 48$)

отметить, что небольшой процент сохранившихся генов как в IG.P1, так и в EVI характерен в основном для маленьких модулей с семью и меньше генами с положительным весом. Среди плохо сохранившихся модулей присутствует также один модуль из 14 генов с положительным весом.

На рис. 5 приведены результаты сравнения референсных модулей с результатами экспериментов с добавлением небольшого шума из нормального распределения. При сравнении с результатами, полученными при экспериментах с варьированием *k*, обнаруживается, что метод менее устойчив к шуму в данных, чем к небольшим отклонениям в числе начальных кластеров. При этом *maxCor* сохраняет высокие значения: все так же только у единичных модулей средняя корреляция меньше 0,7. Метрики же *retainedGenes* и *maxInter* показывают по одному пропавшему модулю в большинстве из экспериментов в каждом из трех наборов данных (число генов с положительным весом в этих модулях: 5, 14, 5).

Таким образом, можно утверждать, что результаты анализа в целом являются достаточно устойчивыми. Для большинства модулей их профили хорошо находятся в шумных данных, сохраняется и большинство генов этих модулей. Однако один из достаточно больших модулей (модуль N8) на наборе данных IG.P1, состоящий из 14 генов с положительным весом) продемонстрировал неустойчивость. Этот эффект требует дальнейшего изучения и возможной корректировки метода.

Заключение

В настоящей работе модифицирован для использования на реальных данных метод совместной кластеризации в графовом и корреляционном пространствах, который позволяет идентифицировать активные модули в наборах транскриптомных данных, представленных большим числом образцов.

В качестве модификаций этого метода предложены способы автоматического подбора параметров метода, обеспечивающие интерпретируемость получаемых результатов. Тестирование метода на реальных данных показало, что большинство полученных им модулей устойчиво к небольшим отклонениям в исходных данных.

Дальнейшие направления разработки предложенного метода могут заключаться в рассмотрении других способов задания весов и параметров, а также пересмотре метрик оценки получаемых результатов. Изучение и валидация полученных модулей другими in silico методами (например, методом анализа баланса потоков) является полезным направлением в интерпретации и подкреплении значимости полученных результатов.

Литература

1. Van den Bossche J., O’Neill L.A., Menon D. Macrophage immunometabolism: where are we (going)? // Trends in Immunology. 2017. V. 38. N 6. P. 395–406. doi: 10.1016/j.it.2017.03.001
2. Al-Khami A.A., Rodriguez P.C., Ochoa A.C. Energy metabolic pathways control the fate and function of myeloid immune cells // Journal of Leukocyte Biology. 2017. V. 102. N 2. P. 369–380. doi: 10.1189/jlb.1VMR1216-535R
3. Wculek S.K., Khouili S.C., Priego E., Heras-Murillo I., Sancho D. Metabolic control of dendritic cell functions: digesting information // Frontiers in immunology. 2019. V. 10. P. 775. doi: 10.3389/fimmu.2019.00775
4. Wang Z., Gerstein M., Snyder M. RNA-Seq: a revolutionary tool for transcriptomics // Nature Reviews Genetics. 2009. V. 10. N 1. P. 57–63. doi: 10.1038/nrg2484
5. Chen G., Ning B., Shi T. Single-cell RNA-Seq technologies and related computational data analysis // Frontiers in Genetics. 2019. V. 10. P. 317. doi: 10.3389/fgene.2019.00317
6. Beisser D., Grohme M.A., Kopka J., Frohme M., Schill R.O., Hengherr S., Dandekar T., Klau G.W., Dittrich M., Müller T. Integrated pathway modules using time-course metabolic profiles and EST data from Milnesium tardigradum // BMC Systems Biology. 2012. V. 6. P. 72. doi: 10.1186/1752-0509-6-72
7. Jha A.K., Huang S.C., Sergushichev A., Lampropoulou V., Ivanova Y., Loginicheva E., Chmielewski K., Stewart K., Ashall J., Everts B., Pearce E., Driggers E.M., Artyomov M.N. Network integration of parallel metabolic and transcriptional data reveals metabolic modules that regulate macrophage polarization // Immunity. 2015. V. 42. N 3. P. 419–430. doi: 10.1016/j.immuni.2015.02.005
8. Artyomov M.N., Sergushichev A., Schilling J.D. Integrating immunometabolism and macrophage diversity // Seminars in immunology. 2016. V. 28. N 5. P. 417–424. doi: 10.1016/j.smim.2016.10.004
9. Sergushichev A.A., Loboda A.A., Jha A.K., Vincent E.E., Driggers E.M., Jones R.G., Pearce E.J., Artyomov M.N. GAM: a web-service for integrated transcriptional and metabolic network analysis // Nucleic acids research. 2016. V. 44. N W1. P. W194–W200. doi: 10.1093/nar/gkw266
10. Гайнуллина А.Н., Шалыто А.А., Сергушичев А.А. Метод совместной кластеризации в графовом и корреляционном простран-

References

1. Van den Bossche J., O’Neill L.A., Menon D. Macrophage immunometabolism: where are we (going)? *Trends in Immunology*, 2017, vol. 38, no. 6, pp. 395–406. doi: 10.1016/j.it.2017.03.001
2. Al-Khami A.A., Rodriguez P.C., Ochoa A.C. Energy metabolic pathways control the fate and function of myeloid immune cells. *Journal of Leukocyte Biology*, 2017, vol. 102, no. 2, pp. 369–380. doi: 10.1189/jlb.1VMR1216-535R
3. Wculek S.K., Khouili S.C., Priego E., Heras-Murillo I., Sancho D. Metabolic control of dendritic cell functions: digesting information. *Frontiers in immunology*, 2019, vol. 10, pp. 775. doi: 10.3389/fimmu.2019.00775
4. Wang Z., Gerstein M., Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 2009, vol. 10, no. 1, pp. 57–63. doi: 10.1038/nrg2484
5. Chen G., Ning B., Shi T. Single-cell RNA-Seq technologies and related computational data analysis. *Frontiers in Genetics*, 2019, vol. 10, pp. 317. doi: 10.3389/fgene.2019.00317
6. Beisser D., Grohme M.A., Kopka J., Frohme M., Schill R.O., Hengherr S., Dandekar T., Klau G.W., Dittrich M., Müller T. Integrated pathway modules using time-course metabolic profiles and EST data from *Milnesium tardigradum*. *BMC Systems Biology*, 2012, vol. 6, pp. 72. doi: 10.1186/1752-0509-6-72
7. Jha A.K., Huang S.C., Sergushichev A., Lampropoulou V., Ivanova Y., Loginicheva E., Chmielewski K., Stewart K., Ashall J., Everts B., Pearce E., Driggers E.M., Artyomov M.N. Network integration of parallel metabolic and transcriptional data reveals metabolic modules that regulate macrophage polarization. *Immunity*, 2015, vol. 42, no. 3, pp. 419–430. doi: 10.1016/j.immuni.2015.02.005
8. Artyomov M.N., Sergushichev A., Schilling J.D. Integrating immunometabolism and macrophage diversity. *Seminars in immunology*, 2016, vol. 28, no. 5, pp. 417–424. doi: 10.1016/j.smim.2016.10.004
9. Sergushichev A.A., Loboda A.A., Jha A.K., Vincent E.E., Driggers E.M., Jones R.G., Pearce E.J., Artyomov M.N. GAM: a web-service for integrated transcriptional and metabolic network analysis. *Nucleic acids research*, 2016, vol. 44, no. W1, pp. W194–W200. doi: 10.1093/nar/gkw266
10. Gainullina A.N., Shalyto A.A., Sergushichev A.A. Method of the joint clustering in network and correlation spaces. *Modeling and Analysis*

ствах // Моделирование и анализ информационных систем. 2020. Т. 27. № 2. С. 180–193. doi: 10.18255/1818-1015-2020-2-180-193
11. Loboda A.A., Artyomov M.N., Sergushichev A.A. Solving generalized maximum-weight connected subgraph problem for network enrichment analysis // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2016. V. 9838. P. 210–221. doi: 10.1007/978-3-319-43681-4_17
12. Benoist C. Open-source ImmGen: Mononuclear phagocytes // Nature Immunology. 2016. V. 17. N 7. P. 741. doi: 10.1038/ni.3478
13. Gautier E.L. Shay T., Miller J., Greter M., Jakubzick C., Ivanov S., Helft J., Chow A., Elpek K.G., Gordonov S., Mazloom A.R., Ma'ayan A., Chua W.-J., Hansen T.H., Turley S.J., Merad M., Randolph G.J., Best A.J., Knell J., Goldrath A., Brown B., Jojic V., Koller D., Cohen N., Brenner M., Regev A., Fletcher A., Bellemare-Pelletier A., Malhotra D., Jianu R., Laidlaw D., Collins J., Narayan K., Sylvia K., Kang J., Gazit R., Garrison B.S., Rossi D.J., Kim F., Rao T.N., Wagers A., Shinton S.A., Hardy R.R., Monach P., Bezman N.A., Sun J.C., Kim C.C., Lanier L.L., Heng T., Kreslavsky T., Painter M., Ericson J., Davis S., Mathis D., Benoist C. Gene-expression profiles and transcriptional regulatory pathways that underlie the identity and diversity of mouse tissue macrophages // Nature Immunology. 2012. V. 13. N 11. P. 1118–1128. doi: 10.1038/ni.2419
14. Röhl S., Rykaczewska U., Seime T., Suur B.E., Diez M.G., Gådin J.R., Gainullina A., Sergushichev A.A., Wirka R., Lengquist M., Kronqvist M., Bergman O., Odeberg J., Lindeman J.H.N., Quertermous T., Hamsten A., Eriksson P., Hedin U., Matic L.P. Transcriptomic profiling of experimental arterial injury reveals new mechanisms and temporal dynamics in vascular healing response // JVS: Vascular Science. 2020. V. 1. P. 13–27. doi: 10.1016/j.jvssci.2020.01.001
15. Kaufman L., Rousseeuw P.J. Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley & Sons, 1990.

of Information Systems, 2020, vol. 27, no. 2, pp. 180–193. (in Russian). doi: 10.18255/1818-1015-2020-2-180-193
11. Loboda A.A., Artyomov M.N., Sergushichev A.A. Solving generalized maximum-weight connected subgraph problem for network enrichment analysis. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2016, vol. 9838, pp. 210–221. doi: 10.1007/978-3-319-43681-4_17
12. Benoist C. Open-source ImmGen: Mononuclear phagocytes. *Nature Immunology*, 2016, vol. 17, no. 7, pp. 741. doi: 10.1038/ni.3478
13. Gautier E.L. Shay T., Miller J., Greter M., Jakubzick C., Ivanov S., Helft J., Chow A., Elpek K.G., Gordonov S., Mazloom A.R., Ma'ayan A., Chua W.-J., Hansen T.H., Turley S.J., Merad M., Randolph G.J., Best A.J., Knell J., Goldrath A., Brown B., Jojic V., Koller D., Cohen N., Brenner M., Regev A., Fletcher A., Bellemare-Pelletier A., Malhotra D., Jianu R., Laidlaw D., Collins J., Narayan K., Sylvia K., Kang J., Gazit R., Garrison B.S., Rossi D.J., Kim F., Rao T.N., Wagers A., Shinton S.A., Hardy R.R., Monach P., Bezman N.A., Sun J.C., Kim C.C., Lanier L.L., Heng T., Kreslavsky T., Painter M., Ericson J., Davis S., Mathis D., Benoist C. Gene-expression profiles and transcriptional regulatory pathways that underlie the identity and diversity of mouse tissue macrophages. *Nature Immunology*, 2012, vol. 13, no. 11, pp. 1118–1128. doi: 10.1038/ni.2419
14. Röhl S., Rykaczewska U., Seime T., Suur B.E., Diez M.G., Gådin J.R., Gainullina A., Sergushichev A.A., Wirka R., Lengquist M., Kronqvist M., Bergman O., Odeberg J., Lindeman J.H.N., Quertermous T., Hamsten A., Eriksson P., Hedin U., Matic L.P. Transcriptomic profiling of experimental arterial injury reveals new mechanisms and temporal dynamics in vascular healing response. *JVS: Vascular Science*, 2020, vol. 1, pp. 13–27. doi: 10.1016/j.jvssci.2020.01.001
15. Kaufman L. Rousseeuw P.J. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, 1990.

Авторы

Гайнуллина Анастасия Наильевна — программист, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, Scopus ID: 57205601752, ORCID: 0000-0003-3796-2337, anastasiia.gainullina@gmail.com

Артемов Максим — PhD, химические науки, профессор (исследователь), Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, Scopus ID: 9242717500, ORCID: 0000-0002-1133-4212, martyomov@pathology.wustl.edu

Сергушичев Алексей Александрович — кандидат технических наук, доцент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, Scopus ID: 55772694000, ORCID: 0000-0003-1159-7220, alserg@itmo.ru

Authors

Anastasiia N. Gainullina — Software Developer, ITMO University, Saint Petersburg, 197101, Russian Federation, Scopus ID: 57205601752, ORCID: 0000-0003-3796-2337, anastasiia.gainullina@gmail.com

Maxim Artyomov — PhD, Chemistry, Professor (Researcher), ITMO University, Saint Petersburg, 197101, Russian Federation, Scopus ID: 9242717500, ORCID: 0000-0002-1133-4212, martyomov@pathology.wustl.edu

Alexey A. Sergushichev — PhD, Associate Professor, Associate Professor, ITMO University, Saint Petersburg, 197101, Russian Federation, Scopus ID: 55772694000, ORCID: 0000-0003-1159-7220, alserg@itmo.ru