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

КРАТКИЕ СООБЩЕНИЯ

BRIEF PAPERS

УДК 519.1 doi: 10.17586/2226-1494-2020-20-6-888-892

ПРИМЕНЕНИЕ МЕТОДА НЕЗАВИСИМЫХ КОМПОНЕНТ ДЛЯ ОПРЕДЕЛЕНИЯ НАЧАЛЬНОГО ПРИБЛИЖЕНИЯ ПРИ ПОИСКЕ АКТИВНЫХ МОДУЛЕЙ В БИОЛОГИЧЕСКИХ ГРАФАХ

А.Н. Гайнуллина, В.Д. Сухов, А.А. Шалыто, А.А. Сергушичев

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

Информация о статье
Поступила в редакцию 23.09.20, принята к печати 30.10.20
Язык статьи — русский

Ссылка для цитирования: Гайнуллина А.Н., Сухов В.Д., Шалыто А.А., Сергушичев А.А. Применение метода независимых компонент для определения начального приближения при поиске активных модулей в биологических графах // Научно-технический вестник информационных технологий, механики и оптики. 2020. Т. 20. № 6. С. 888–892. doi: 10.17586/2226-1494-2020-20-6-888-892

Аннотация

Предмет исследования. Поиск активных модулей в биологических графах, в том числе в генных графах, является одним из важных подходов к интерпретации экспериментальных биологических данных. Один из методов ее решения основан на применении алгоритма совместной кластеризации в графовом и корреляционном пространствах. Алгоритм находит группы генов, одновременно близко расположенные в генном графе и обладающие высокой попарной корреляцией по матрице значений экспрессии генов. Алгоритм является итеративным, один из его ключевых параметров — выбранное начальное приближение, от которого зависит время работы и качество получаемых результатов. В настоящей работе рассмотрена задача определения начального приближения для этого алгоритма. Для решения задачи предложено использование процедуры на основе метода независимых компонент. **Метод.** На первом шаге предлагаемой процедуры определения начального приближения применяется метод независимых компонент к центрированной матрице значений экспрессии генов. Далее для каждой компоненты определяются гены, которые ей соответствуют с заданным уровнем статистической значимости. Полученные группы генов для всех независимых компонент выбираются в качестве начального приближения. **Основные результаты.** Применение процедуры на основе метода независимых компонент позволит уменьшить число групп генов в начальном приближении без потери точности, что, в свою очередь, уменьшит время работы алгоритма кластеризации в десятки раз при сохранении качества результатов. **Практическая значимость.** Ускорение работы алгоритма совместной кластеризации в графовом и корреляционном пространствах без потери качества результатов значительно повысит удобство его использования для интерпретации транскриптомных данных в биоинформатике и вычислительной биологии.

Ключевые слова
кластеризация, корреляция, метод независимых компонент, графы, экспрессия генов

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

doi: 10.17586/2226-1494-2020-20-6-888-892

INDEPENDENT COMPONENT ANALYSIS FOR INITIAL APPROXIMATION DETERMINATION IN IDENTIFICATION OF ACTIVE MODULES IN BIOLOGICAL GRAPHS

A.N. Gainullina, V.D. Sukhov, A.A. Shalyto, A.A. Sergushichev

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

Article info
Received 23.09.20, accepted 30.10.20
Article in Russian

For citation: Gainullina A.N., Sukhov V.D., Shalyto A.A., Sergushichev A.A. Independent component analysis for initial approximation determination in identification of active modules in biological graphs. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2020, vol. 20, no. 6, pp. 888–892 (in Russian). doi: 10.17586/2226-1494-2020-20-6-888-892

Abstract

Subject of Research. The identification of active modules in biological graphs, for example, gene graphs, is one of the important approaches to the interpretation of experimental biological data. One of the approaches for its solution is the application of an algorithm of the joint clustering in network and correlation spaces. The algorithm finds groups of genes that are located simultaneously close in the gene graph and have a high pairwise correlation according to the matrix of gene expression values. The algorithm is iterative and one of its key parameters is the chosen initial approximation, which affects both the run time and the quality of the results. We consider the determination problem of an initial approximation for this algorithm. A procedure based on independent component analysis is proposed for the problem solution. **Method.** The method of independent component analysis is applied to a centered matrix of expression values at the first step of the proposed procedure for finding of an initial approximation. Then, the genes specific to the component with a given level of statistical significance are identified for each component. The gene groups obtained for all independent components are chosen as the initial approximation. **Main Results.** The procedure application based on the independent component analysis reduces the number of gene groups in the initial approximation without the loss of accuracy. This fact, in turn, speeds up the running time of the clustering algorithm by an order of magnitude with the quality maintenance of the results. **Practical Relevance.** Acceleration of the algorithm of the joint clustering in network and correlation spaces without quality loss of the results increases significantly its convenience and simplifies its application for the interpretation of transcriptome data in bioinformatics and computational biology.

Keywords

clustering, correlation, independent component analysis, graphs, gene expression

Acknowledgements

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

Задача поиска активных модулей в биологических графах возникает при интерпретации экспериментальных данных, полученных при изучении некоторых биологических процессов [1–3]. В рамках этого подхода рассматривается граф взаимодействий некоторых биологических элементов, например, генов, и предполагается, что в этом графе существуют связные подграфы — *активные модули*, регуляция которых происходит совместно [4]. В случае генов такая совместная регуляция может выражаться в скоррелированных уровнях экспрессии (или активности) генов, что можно наблюдать экспериментально с помощью метода РНК-секвенирование. Задача поиска активных модулей в этом случае состоит в том, чтобы по данным об экспрессии генов определить потенциальные активные модули.

Для решения этой задачи был предложен алгоритм совместной кластеризации в графовом и корреляционном пространствах *net-clust* [5]. Входными данными алгоритма служат матрица со значениями экспрессии генов в некоторых биологических образцах и граф генных взаимодействий. На выходе алгоритм выдает набор связных подграфов, таких, что профили экспрессии генов сильно коррелируют внутри каждого подграфа. На начальном шаге определяются группы хорошо скорреллированных генов без требования к их связности (*начальное приближение*), и по ним определяются потенциальные профили экспрессии активных модулей. Затем для каждого профиля находится связный подграф, гены которого хорошо коррелируют с рассматриваемым профилем. По найденным связным подграфам корректируются потенциальные профили экспрессии активных модулей, далее эти шаги повторяются до тех пор, пока и модули, и потенциальные профили не перестанут изменяться.

Важным элементом описанного алгоритма является процедура определения начального приближения. В статье [5] для этого рассматривались два метода кластеризации: *k-means* и *k-medoids* с разными значениями параметра k. При этом на симулированных данных было показано, что для качественной работы алгоритма значения k должны быть в несколько раз больше истинного числа модулей. С другой стороны, увеличение значения k влечет нелинейное увеличение числа итераций и времени работы алгоритма.

В настоящей работе рассматривается проблема получения начального приближения для алгоритма совместной кластеризации в графовом и корреляционном пространствах, и для ее решения предлагается использовать процедуру, основанную на применении метода независимых компонент (Independent Component Analysis, ICA) [6–8].

Предлагаемая процедура основана на модели, в которой матрицу $\mathbf{E}$ значений экспрессии генов можно представить в виде

$$\mathbf{E} = \mathbf{S} \times \mathbf{A} + \boldsymbol{\varepsilon},$$

где $\mathbf{A}$ — матрица смеси, соответствующая тому, какие активные модули представлены в каких образцах; $\mathbf{S}$ — матрица сигналов, соответствующая тому, какие гены и с каким весом входят в активные модули; $\boldsymbol{\varepsilon}$ — матрица шума, соответствующая как техническим, так и биологическим случайным эффектам.

В рамках этой модели и некоторых других предположений метод ICA позволяет по матрице $\mathbf{E}$ получить разложение на матрицы $\mathbf{S}$ и $\mathbf{A}$.

Важным свойством этого разложения является то, что значения в каждом столбце матрицы $\mathbf{S}$ (соответствуют некоторой независимой компоненте) распределены согласно нормальному распределению, кроме «тяжелых хвостов». «Хвосты» соответствуют генам, для которых можно с уверенностью сказать, что они принадлежат этой компоненте. При этом уверенность можно контролировать с помощью задания порога на долю ложноположительных срабатываний (False Discovery Rate, FDR).

Таким образом для каждой независимой компоненты можно определить две группы генов, принадлежащих этой компоненте: гены с большими положительными значениями в соответствующей колонке матрицы $\mathbf{S}$,

которые изменяются в том же направлении, что и компонента, и гены с большими отрицательными значениями, которые изменяются в противоположном направлении. Убрав гены, входящие в несколько компонент, и убрав группы с небольшим числом генов, получим набор групп генов с хорошей внутренней попарной корреляцией. Эти группы и будут использоваться в качестве начального приближения для алгоритма *net-clust*.

Параметром метода ICA является число получаемых компонент. Для выбора подходящего числа компонент существуют различные подходы [9, 10]. В настоящей работе для простоты используется следующий подход. Для исходной матрицы **E** запускается метод главных компонент (Principal Component Analysis, PCA). Для полученных компонент вычисляется доля вариации в исходных данных, объясняемая каждой из этих компонент. Наконец, в качестве числа компонент для запуска метода ICA выбирается число компонент, полученных методом PCA, для которых доля объясняемой вариации превышает 5 %.

Для экспериментального исследования алгоритма с получением начального приближения на основе метода ICA сгенерированы симулированные данные, согласно описанному в [5] протоколу. Было рассмотрено три типа комбинаций истинных модулей в данных экспрессии, соответствующих различным дизайнам биологических экспериментов. Для всех типов рассматривалось шесть биологических состояний и каждое состояние было представлено в трех повторностях — типичном числе для биологических экспериментов, в которых анализируется экспрессия генов.

Сначала качество результатов работы алгоритма *net-clust* при построении начального приближения на основе метода ICA сравнивалось с качеством результатов, полученных при использовании методов *k-medoids* и *k-means* для $k = 32$, предлагавшихся к использованию в [5]. При сравнении варьировалось два параметра: параметр σ — значение среднеквадратичного отклонения шума, использующегося при генерации симулированных данных, и параметр *base* — параметр алгоритма *net-clust*, контролирующий порог корреляции для включения генов в модуль (порог корреляции вычисляется как 1-*base*) и тем самым контролирующий соотношения метрик точности и полноты получаемых результатов. В этом эксперименте всегда генерировалось десять истинных модулей.

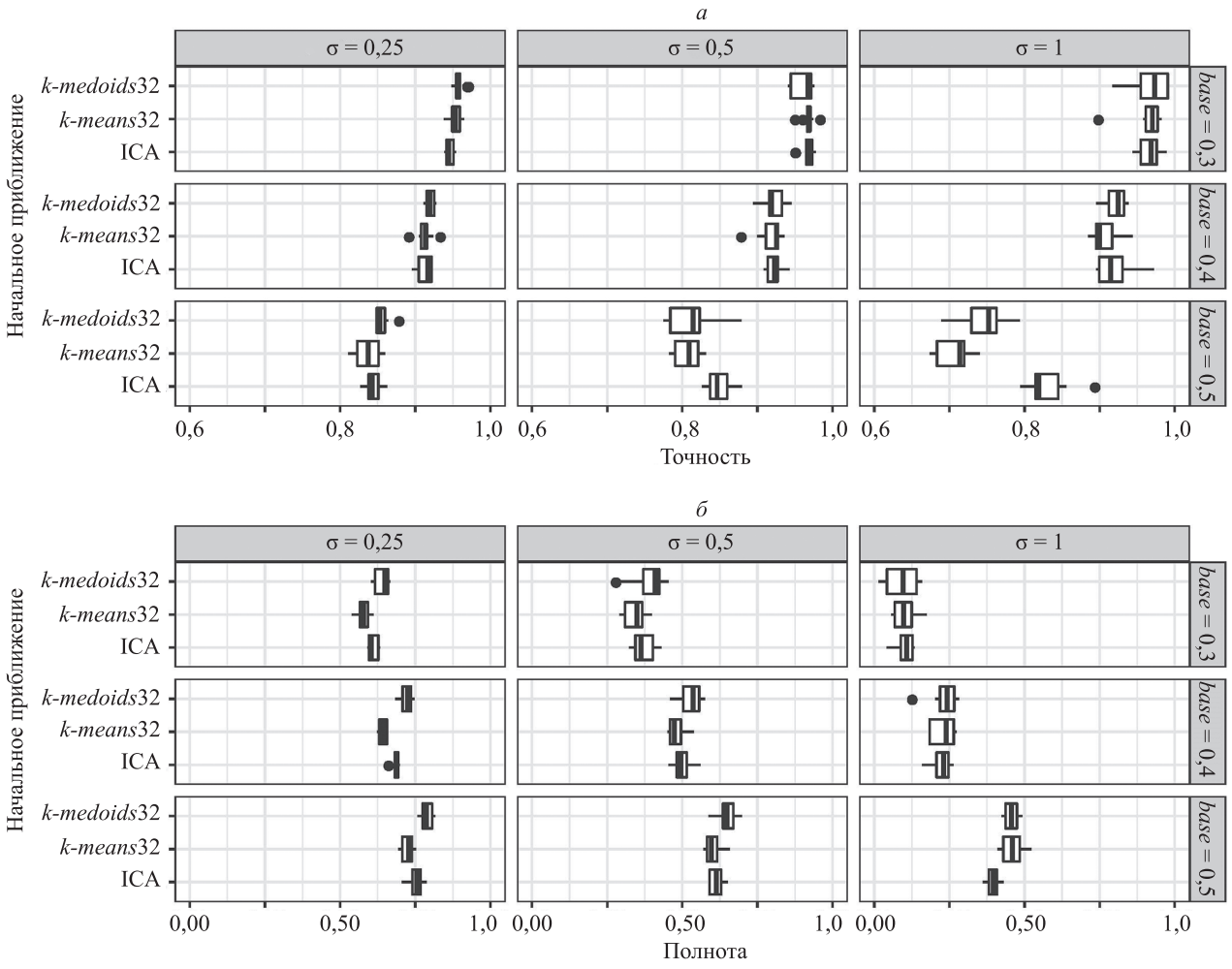

Рис. 1. Анализ результатов алгоритма *net-clust* при разных способах получения начального приближения: *k-means*, *k-medoids* и ICA: точность (*а*); полнота (*б*).

σ – значение среднеквадратичного отклонения шума, использующегося при генерации симулированных данных; *base* — параметр алгоритма *net-clust*, контролирующий порог корреляции для включения генов в модуль

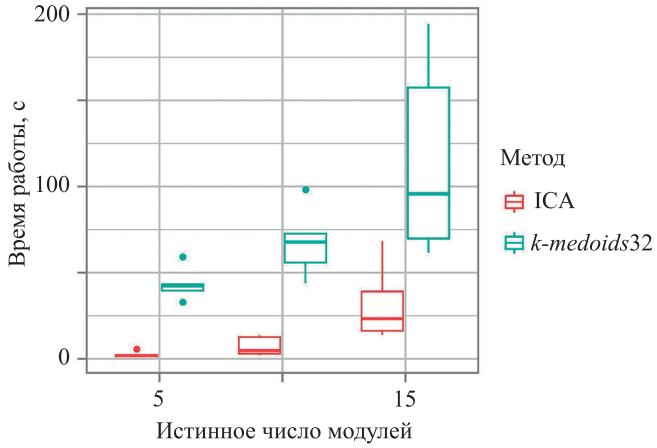

Рис. 2. Время работы алгоритма *net-clust* при использовании одного из методов получения начального приближения (*k-medoids* при $k = 32$ или ICA); истинное число активных модулей в симулированных данных равнялось 5, 10 или 15; значение среднеквадратичного отклонения для шума равнялось 0,5; значение параметра *base* равнялось 0,4

На рис. 1 приведены результаты этого сравнения. Можно наблюдать, что качество результатов при получении начального приближения с помощью метода ICA достаточно похоже на качество результатов при использовании методов *k-means* и *k-medoids*. Исключение составляет результат при $\sigma = 1$ и *base* = 0,5, для которых использование метода ICA дает большую точность, но меньшую полноту.

Выполнен анализ времени работы алгоритма *net-clust* в зависимости от метода построения начального приближения. В сравнении участвовали методы *k-medoids* при $k = 32$ и ICA. В этом эксперименте значение среднеквадратичного отклонения для шума σ приравнивалось 0,5, как типичному значению для реальных экспериментальных данных [11]. При этом варьировалось число истинных модулей (5, 10 или 15).

На рис. 2 представлены результаты анализа времени работы. Применение процедуры на основе ICA уменьшает общее время работы алгоритма по сравнению с методом *k-medoids* в 25, 14 и 4 раза для случаев, когда число истинных модулей равно 5, 10 и 15 соответственно. Это объясняется тем, что процедура на основе ICA возвращает значительно меньшее число кластеров, примерно совпадающее с истинным значением числа модулей, тем самым уменьшая число итераций алгоритма *net-clust*.

Таким образом, в настоящей работе предложено использование метода независимых компонент для получения начального приближения в алгоритме совместной кластеризации в графовом и корреляционном пространствах *net-clust*. За счет меньшего числа групп генов в получаемом приближении этот подход позволяет значительно уменьшить время работы алгоритма *net-clust* при сохранении качества получаемых результатов.

Литература

1. 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
2. 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
3. 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
4. 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
5. Гайнуллина А.Н., Шалыто А.А., Сергушичев А.А. Метод совместной кластеризации в графовом и корреляционном пространствах // Моделирование и анализ информационных систем. 2020. Т. 27. № 2. С. 180–193. doi: 10.18255/1818-1015-2020-2-180-193
6. Comon P. Independent component analysis, a new concept? // Signal Processing. 1994. V. 36. N 3. P. 287–314. doi: 10.1016/0165-1684(94)90029-9
7. Saelens W., Cannoodt R., Saeys Y. A comprehensive evaluation of module detection methods for gene expression data // Nature Communications. 2018. V. 9. N 1. P. 1090. doi: 10.1038/s41467-018-03424-4
8. Rotival M., Zeller T., Wild P., Maouche S., Szymczak S., Schillert A., Castagné R., Deiseroth A., Proust C., Brocheton J., Godefroy T., Perret C., Germain M., Eleftheriadis M., Sinning C.R., Schnabel R.B., Lubos E., Lackner K.J., Rossmann H., Münzel T., Rendon A., Consortium C., Erdmann J., Deloukas P., Hengstenberg C., Diemert P., Montalescot G., Ouwehand W.H., Samani N.J.,

References

1. 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
2. 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
3. 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
4. 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
5. Gainullina A.N., Shalyto A.A., Sergushichev A.A. Method of the joint clustering in network and correlation spaces. *Modeling and Analysis of Information Systems*, 2020, vol. 27, no. 2, pp. 180–193. (in Russian). doi: 10.18255/1818-1015-2020-2-180-193
6. Comon P. Independent component analysis, a new concept? *Signal Processing*, 1994, vol. 36, no. 3, pp. 287–314. doi: 10.1016/0165-1684(94)90029-9
7. Saelens W., Cannoodt R., Saeys Y. A comprehensive evaluation of module detection methods for gene expression data. *Nature Communications*, 2018, vol. 9, no. 1, pp. 1090. doi: 10.1038/s41467-018-03424-4
8. Rotival M., Zeller T., Wild P., Maouche S., Szymczak S., Schillert A., Castagné R., Deiseroth A., Proust C., Brocheton J., Godefroy T., Perret C., Germain M., Eleftheriadis M., Sinning C.R., Schnabel R.B., Lubos E., Lackner K.J., Rossmann H., Münzel T., Rendon A., Consortium C., Erdmann J., Deloukas P., Hengstenberg C., Diemert P., Montalescot G., Ouwehand W.H., Samani N.J.,

Schunkert H., Tregouet D.-A., Ziegler A., Goodall A.H., Cambien F., Tiret L., Blankenberg S. Integrating genome-wide genetic variations and monocyte expression data reveals trans-regulated gene modules in humans // PLoS Genetics. 2011. V. 7. N 12. P. e1002367. doi: 10.1371/journal.pgen.1002367

9. Minka T. Automatic choice of dimensionality for PCA // Advances in Neural Information Processing Systems. 2001. V. 13. P. 598–604.
10. Ray K.L., McKay D.R., Fox P.M., Riedel M.C., Uecker A.M., Beckmann C.F., Smith S.M., Fox P.T., Laird A.R. ICA model order selection of task co-activation networks // Frontiers in Neuroscience. 2013. V. 7. P. 237. doi: 10.3389/fnins.2013.00237
11. Steinbaugh M.J., Pantano L., Kirchner R.D., Barrera V., Chapman B.A., Piper M.E., Mistry M., Khetani R.S., Rutherford K.D., Hofmann O., Hutchinson J.N., Sui S.H. BcbioRNASeq: R package for bcbio RNA-seq analysis // F1000Research. 2017. V. 6. P. 1976. doi: 10.12688/f1000research.12093.1

Schunkert H., Tregouet D.-A., Ziegler A., Goodall A.H., Cambien F., Tiret L., Blankenberg S. Integrating genome-wide genetic variations and monocyte expression data reveals trans-regulated gene modules in humans. *PLoS Genetics*, 2011, vol. 7, no. 12, pp. e1002367. doi: 10.1371/journal.pgen.1002367

9. Minka T. Automatic choice of dimensionality for PCA. *Advances in Neural Information Processing Systems*, 2001, vol. 13, pp. 598–604.
10. Ray K.L., McKay D.R., Fox P.M., Riedel M.C., Uecker A.M., Beckmann C.F., Smith S.M., Fox P.T., Laird A.R. ICA model order selection of task co-activation networks. *Frontiers in Neuroscience*, 2013, vol. 7, pp. 237. doi: 10.3389/fnins.2013.00237
11. Steinbaugh M.J., Pantano L., Kirchner R.D., Barrera V., Chapman B.A., Piper M.E., Mistry M., Khetani R.S., Rutherford K.D., Hofmann O., Hutchinson J.N., Sui S.H. BcbioRNASeq: R package for bcbio RNA-seq analysis. *F1000Research*, 2017, vol. 6, pp. 1976. doi: 10.12688/f1000research.12093.1

Авторы

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

Сухов Владимир Дмитриевич — программист, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, ORCID: 0000-0002-5169-1433, vdsukhov@yandex.ru

Шалыто Анатолий Абрамович — доктор технических наук, профессор, главный научный сотрудник, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, Scopus ID: 56131789500, ORCID: 0000-0002-2723-2077, shalyto@mail.ifmo.ru

Сергушичев Алексей Александрович — кандидат технических наук, доцент, 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

Vladimir D. Sukhov — Software Developer, ITMO University, Saint Petersburg, 197101, Russian Federation, ORCID: 0000-0002-5169-1433, vdsukhov@yandex.ru

Anatoly A. Shalyto — D.Sc., Professor, Chief Researcher, ITMO University, Saint Petersburg, 197101, Russian Federation, Scopus ID: 56131789500, ORCID: 0000-0002-2723-2077, shalyto@mail.ifmo.ru

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