УДК 519.2:004.421.5:004.7
DOI: 10.25206/1813-8225-2018-162-265-270

М. Н. ЮДИНА

Омский государственный
технический университет,
г. Омск

КОМПЛЕКС ПРОГРАММНЫХ БИБЛИОТЕК ДЛЯ АНАЛИЗА МОЛЕКУЛЯРНЫХ СЕТЕЙ КЛЕТКИ

Предлагается комплекс программных библиотек, предназначенный для анализа молекулярных сетей клетки. В работе рассматриваются возможности разработанного комплекса для исследования сетей белковых взаимодействий и генных сетей. Описываются решаемые программным комплексом задачи и характеризуются оригинальные алгоритмы, реализованные в разработанном комплексе программных библиотек.

Ключевые слова: молекулярные сети клетки, пакеты программ для исследования сетей, структурные характеристики больших сетей, значимые типовые подграфы, случайные графы с нелинейным правилом предпочтительного связывания, ускоренные алгоритмы и численные методы.

Введение. Анализ молекулярных сетей клетки является драйвером развития биоинформатики в последние десятилетия: такой анализ помогает лучше понять структуру взаимодействия генов, белков, метаболитов и других химико-биологических составляющих клетки. С системной точки зрения, особый интерес вызывают случаи, когда молекулярные сети клетки имеют схожие структурные характеристики как между собой, так и в сравнении с другими большими сетями, такими как социальные сети, сети телекоммуникаций, сети соавторства и т.д. Причем эти общие характеристики изучаются как в рамках самой биоинформатики, так и в рамках другой междисциплинарной науки, появившейся на пороге XXI века — Науки о сетях (Network Science).

В частности, к таким общим характеристикам многих сетей относятся [1, 2]:

— небольшое среднее расстояние между узлами (например, для многих сетей выполняется правило шести рукопожатий);

— высокий средний коэффициент кластеризации (коэффициент кластеризации узла определяет вероятность связей между собой ближайших «соседей» узла);

— степенной закон распределения локальной степени связности узлов (если бы ребра были распределены равновероятно между любыми парами вершин, то распределение было бы пуассоновское).

Каждое из этих свойств повлияло на представление о сетях и понимание процессов, которые в них протекают. Так, наличие степенного закона распределения степени связности узлов в сети приводит к нулевому порогу распространения вирусов [1] и необходимости выработки новых стратегий вакцинации [3]. Наличие степенного закона также влечет такие свойства сетей как устойчивость к случайным удалениям связей и узлов и большую чувствительность к удалению узлов с большим числом связей. Для молекулярных сетей клетки это можно проиллюстрировать наблюдением, представленным в работе [4]: до 73 % генов в клетке S. Cerevisiae (пекарских дрожжах) несущественны, т.е. их удаление не имеет фенотипических эффектов. Это подтверждает надежность генных сетей к случайным сбоям. В то же время вероятность того, что ген является существенным (летальным) или чувствительным к токсинам, зависит от числа взаимодействий, которое имеет его белковый продукт [5]. Типичным примером является ген-супрессор опухоли TP53 и его белковый продукт p53, который является фактором транскрипции и участвует во многих процессах (как правило, регулирующих клеточный цикл).

В связи с большим значением в биоинформатике комплексного исследования аспекта сетевых взаимодействий молекул внутри клетки в последние годы появились различные программы и веб-сервисы, позволяющие исследовать молекулярные сети клетки. Так, популярным веб-ресурсом для анализа генных сетей является веб-ресурс cBioPortal.org, который делает доступным анализ регуляторных сетей посредством использования данных международного проекта, получившего название «Атлас ракового генома» (The Cancer Genome Atlas — TCGA). На рис. 1 представлен снимок экрана с веб-ресурса cBioPortal.org при исследовании сети, описывающей взаимодействие генов, связанных с появлением рака молочной железы по результатам исследования [6]. В частности, с помощью интерфейса приложения выделены ген DIC3CA, а также описанный ранее ген TP53. Ген DIC3CA, в свою очередь, кодирует белок фосфоинозитид-3-киназу, являющуюся ключевым элементом сигнального пути, характерного для большинства клеток человека, и ответственную, в том числе, за уход от апоптоза (процесса программируемой клеточной гибели). Причем даже при анализе взаимодействий только этих двух генов задействовано 712 других (на рис. 1 представлены только 50 из них). Такое большое число задействованных генов — типичная ситуация при исследовании генных сетей.

Помимо многочисленных веб-сервисов для анализа молекулярных сетей клетки существует множество программных пакетов и программ для проведения соответствующих исследований. К таким программам относятся как программы общего

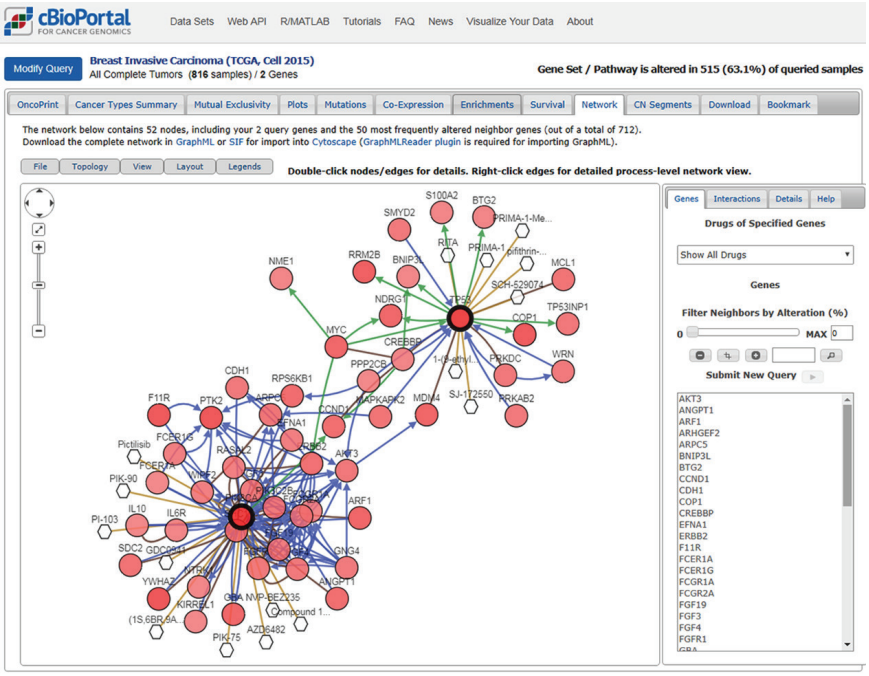

Рис. 1. Генная сеть. Изображение получено онлайн на веб-ресурсе http://www.cbioportal.org/

Таблица 1

Основные характеристики биологических молекулярных сетей

Название характеристики	PPI_HS_U	YeastS	GenReg	Regulon	Pathway Commons
Дата размещения на платформе CyNDEx 2	03.10. 2018	Данные из работы [12]	07.12.17	16.06.2018	22.08. 2017
Тип графа	Неориентир.	Неориентир.	Ориентир.	Ориентир.	Ориентир.
Вершин	22036	2361	16859	21097	19987
Ребер/(дуг)	309245	6646	713649	547408	824675
Число компонентов связности	27	101	1	1	1
Средний коэффициент кластеризации	0,1298639	0,2001346	0,08712398	0,02858035	0,2818434
Средняя длина кратчайших путей	3,229954	4,376182	3,179365	2,966	2,965538

назначения, используемые и в других науках, связанных с Теорией сетей (например Pajek, Gephi, igraph для системы R), так и специальные программы, созданные для специалистов в области биоинформатики (Cytoscape, BioFabric, GeneNet для системы R).

В данной работе предлагаются три программные библиотеки общего назначения, которые могут быть также полезны специалистам, изучающим молекулярные сети клетки. Преимуществом разработанных библиотек является учет характеристик, которые во многих пакетах не учитываются, наличие оригинальных моделей сетей на основе случайных графов, а также ускоренных алгоритмов расчета структурных характеристик с использованием параллельных вычислений.

1. Анализируемые сети. В качестве базовых сетей для демонстрации возможностей разработанных библиотек рассмотрим следующие пять молекулярных сетей клетки, представленных в табл. 1:

1) сеть белковых взаимодействий PPI_HS_U, описывающая взаимодействия белков в клетках человека, по данным проекта BioGRID (https://thebiogrid.org);

2) сеть белковых взаимодействий YeastS, построенная на основе данных о взаимодействии белков в клетке дрожжей S. Cerevisiae [7];

3) регуляторная сеть GenReg, описывающая взаимодействие генов, задействованных в появлении рака коры надпочечника у человека [8];

4) генная сеть Regulon, содержащая информацию о генах, локализованных в различных местах одного генома, но образующая общий механизм экспрессии и связанных с появлением рака коры надпочечника у человека [9];

5) генная сеть PathwayCommons, содержащая информацию о влиянии различных генов на активацию генетических болезней [10].

Одним из способов доступа к данным о молекулярных сетях клетки является использование программы CyNDEx 2, которая встроена в систему Cytoscape и позволяет импортировать данные из базы данных NDEx (Network Data Exchange [11]). В табл. 1 представлены глобальные структурные характеристики сетей, исследуемых в данной работе.

2. Библиотека классов SocAndBioNetworksAnalysis. Предлагаемая в статье библиотека классов SocAndBioNetworksAnalysis представляет собой

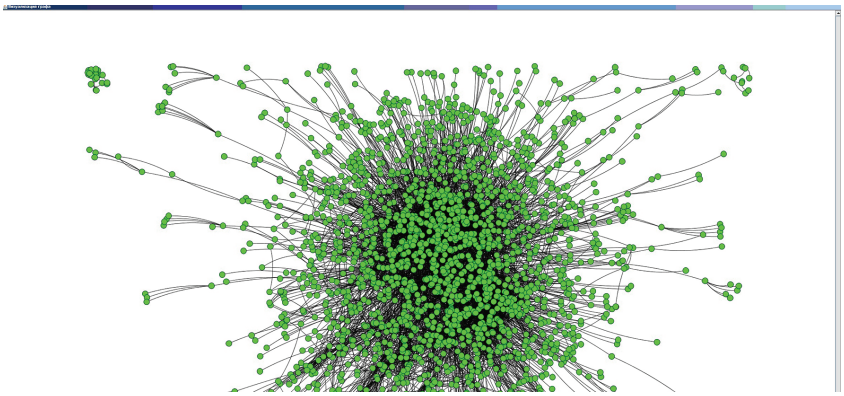

Рис. 2. Визуализации сети YeastS средствами Библиотеки классов SocAndBioNetworksAnalysis

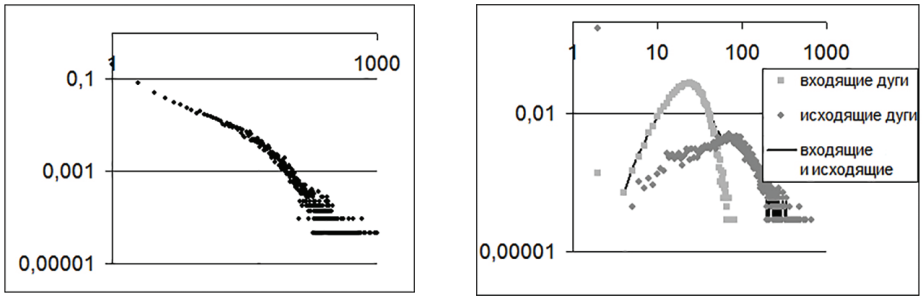

Рис. 3. Распределение степени вершин для сети PPI_HS_U (слева) и Regulon (справа)

Таблица 2

Результаты анализа важности узлов в биологических сетях клетки

Название сети	Используемые метрики важности узлов	
	Важность по числу связей	Важность по промежуточности
PPI_HS_U	TRIM25	TRIM25
YeastS	YPR110C	YNL189W
GenReg	RNF14	MSF
Regulon	FHL1	ELL2
PathwayCommons	SP1	MYC

программный код на языке Java, реализующий алгоритмы анализа биологических и социальных сетей, которая обеспечивает следующую функциональность:

— сохранение и загрузка графа в формате списка ребер, pajek, GraphML;

— визуализация графа (рис. 2);

— расчет ряда структурных характеристик, в том числе расчет распределения степени связности вершин (рис. 3), а также двухмерного распределения степени связности ребер графа (рис. 4)(важность этой характеристики неоднократно подчеркивалась при моделировании больших сетей [13, 14]);

— преобразование графа, представленного в формате сетей Стенфордского университета (http://snap.stanford.edu/data/), в формат типа «список ребер»;

— определение важности узлов с учетом различных мер важности узлов (удаленность до других узлов, степень связности, степень посредничества и др.).

Наиболее важные узлы рассмотренных молекулярных сетей представлены в табл. 2. В качестве мер важности узлов использовались суммарное число связей с другими узлами (*число связей*) и то, как часто узел находится на кратчайших путях между всеми другими узлами попарно (*промежуточность*).

Ниже приведено описание представленных в табл. 2 узлов, описание выполнено по резюме генов и результатов экспрессии генов, представленным на следующих веб-ресурсах: https://www.genecards.org, https://www.uniprot.org, https://www.ncbi.nlm.nih.gov.

1. В сети PPI_HS_U по выбранным показателям особое значение имеет белок TRIM25, который локализуется в цитоплазме. Этот белок может действовать как фактор транскрипции.

2. В сети YeastS по выбранным данным показателям наиболее важным является белок YPR110C — субъединица AC40, которая является общей для РНК-полимеразы I и III. Другим важным белком

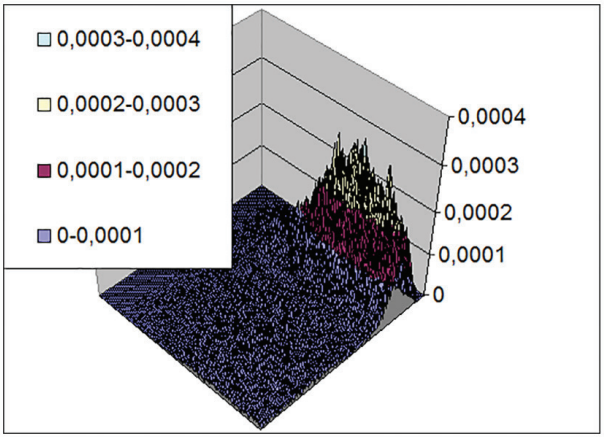

Рис. 4. Двухмерное распределение концевых степеней ребер для сети Regulon

является белок YNL189W, который, помимо прочего, участвует в котрансляционной деградации белка и связывается с рибосомой зарождающейся полипептиды.

3. В сети GenReg наиболее важным является ген RNF14, который кодирует фермент RNF14, содержащий т.н. цинковый палец — белковый модуль, взаимодействующий с ДНК, РНК и другими белками или небольшими молекулами, т.е. активно участвующий в белок-белковых взаимодействиях. Другой важный ген — MSF, который является членом семейства септинов, участвующих в цитокинезе и контроле клеточного цикла.

4. В сети Regulon наиболее важным является ген ELL2, который ответствен за фактор удлинения транскрипции RNA polymerase II.

5. В сети PathwayCommons наиболее важными являются ген SP1, ответственный за экспрессию белка Sp1, который напрямую связывается с ДНК и усиливает уровень транскрипции других генов, а также ген MYC, ответственный за протоонкогенный белок Myc. Регуляция Myc нарушена в 70 % случаев рака [15], т.е. этот белок является очень привлекательной мишенью для противораковой терапии.

3. Библиотека классов MotifsLib. Исследуя молекулярные сети клетки, можно заметить, что некоторые подграфы встречаются чаще, чем это было бы в случайном графе. Некоторые сети состоят из трех-четырех типовых подграфов, в то время как другие типовые подграфы просто не встречаются. Это наблюдение привело к пониманию того, что некоторые типовые подграфы играют важную роль в функционировании сети. Например, в генетических сетях таким типовым подграфом является подграф, получивший название «Feed-forward loop» (рис. 5), когда один ген регулирует другой напрямую и путем регулирования другого гена [16].

Большинство программ для расчета значимых типовых подграфов таких программ, как MFinder [17] (2003), Mavisto [18] (2005), Fanmod [19] (2006), NeMoFinder [20] (2006) и более поздняя реализации того же функционала в программе LaMoFinder, Kavash [21] (2009), библиотека igraph для системы R (2013), AccMotif [22], (2013 год), не позволяют рассчитывать частоты появления подграфов в больших сетях (больше десятков тысяч узлов) за приемлемое время. Целью большинства программ является ускорение расчета встречаемости подграфов на большем числе узлов (чем три и четыре вершины). При этом полагается, что сети содержат не больше нескольких тысяч узлов и связей. Так, в перечисленных выше работах [17 — 22], в которых предлагаются новые алгоритмы для расчета частот встречаемости подграфов на четырех вершинах, наибольшей исследуемой сетью, рассмотренной в этих статьях, является сеть Foldoc. При этом сеть Foldoc содержит всего 12905 узлов и 109092 связи. Сеть описывает связи между терминами онлайн-библиотеки http://www.foldoc.org/. Для ее расчета наиболее быстрым программам требуется 559 секунд (Kavash), 580 секунд (Fanmod), 18 секунд (igraph).

Для больших сетей (а с каждым годом становятся известны данные о все большем количестве взаимодействий между молекулами клетки) для расчета типовых подграфов требуется огромное время расчетов. Так, в табл. 3 приведены оценки времени расчета исследуемых нами сетей на компьютере с четырехъядерным процессором Intel Xeon E3-1245 3,3 ГГц и объемом ОЗУ 8 Гб. Заметим также, что время для определения значимости подграфов в сотни раз больше указанного в таблице, поскольку подразумевает необходимость генерации случайных графов и расчета встречаемости подграфов для них.

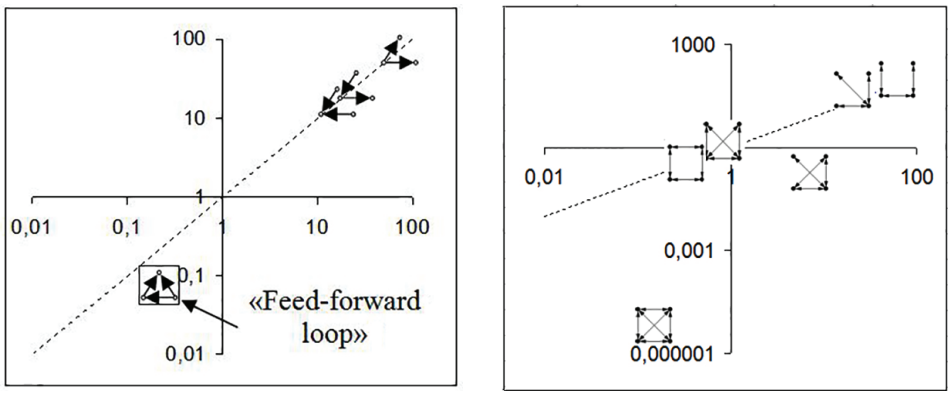

Рис. 5. Слева представлены соотношения частот встречаемости некоторых типовых подграфов на трех вершинах для сети регуляции генов Regulon, справа — типовые подграфы на четырех вершинах для сети белковых взаимодействий Yeast, по оси абсцисс — процент встречаемости типового подграфа в случайном графе, по оси ординат — процент встречаемости в исследуемой сети. Пунктиром изображена функция y = x, типовые подграфы, отображенные ниже пунктирной линии, чаще встречаются в исследуемой сети, чем в случайных графах

Таблица 3

Время расчета типовых подграфов с использованием среды R и программы igraph для исследуемых молекулярных сетей клетки

Сеть	Время расчета типовых подграфов в секундах			
	igraph для среды R		Fanmod	
	на 3-х узлах	на 4-х узлах	на 3-х узлах	на 4-х узлах
PPI_HS_U	35	18824	381,026	182161
YeastS	0,01	0,35	0,052	1,62
GenReg	44	15163	513,986	188820
Regulon	13	2367	92,173	13361,8
PathwayCommons	300	388056	323,9	$>10^4$

Поэтому нами разработаны [23, 24] методы, алгоритмы и библиотека классов MotifsLib для ускорения расчета частот встречаемости подграфов за счет 1) распараллеливания расчета и 2) реализации метода статистического моделирования (с контролем точности).

Теоретическое обоснование и доказательство эффективности используемого нами подхода для расчета частот встречаемости подграфов в больших сетях, по сравнению с другими известными подходами [17–22], можно найти в процитированных выше работах [23, 24].

4. Библиотека классов YMN_GraphGenerators. В отличие от описанных библиотек классов SocAndBioNetworksAnalysis и MotifsLib, библиотека классов **YMN_GraphGenerators** решает не задачу анализа сетей, а задачу их моделирования на основе случайных графов предпочтительного связывания. Модели сетей позволяют прогнозировать динамику изменений в сети, а также выдвигать гипотезы о движущих силах, приводящих к этой динамике.

В частности, в библиотеке реализован оригинальный генератор графов предпочтительного связывания с добавлением стохастических приращений, для которого в работе [25] выведены формулы подбора параметров генерации графа с учетом функции распределения и коэффициента кластеризации.

С помощью представленного генератора, используя эмпирические данные о моделируемых сетях, можно генерировать случайные графы с заданным распределением степени связности вершин, с заданным распределением степени связности концевых вершин дуг, управлять значением коэффициента кластеризации.

Заключение. Разработанные библиотеки классов могут быть использованы для анализа и моделирования молекулярных сетей клетки. Условия использования разработанных автором библиотек подчиняются лицензии GNU Lesser GPL v2.1, благодаря чему сторонние разработчики могут использовать классы библиотеки. Исходные классы библиотек размещены в сети Интернет по адресу: https://github.com/MNYudina. Также эти библиотеки зарегистрированы в Объединенном фонде электронных ресурсов «Наука и образование» [26–28].

Автор выражает благодарность своему научному руководителю — В. Н. Задорожному за ценные советы и обсуждения при разработке описанных в данной статье библиотек классов.

Библиографический список

1. Barabasi A., Bonabeau E. Scale-Free Networks // Scientific American. 2003. Vol. 288. P. 60–69. DOI: 10.1038/scientificamerican0503-60.

2. Albert R., Barabasi A. Statistical mechanics of complex networks // Reviews of Modern Physics. 2002. Vol. 74. P. 47–97. DOI: 10.1103/RevModPhys.74.47.

3. Chakrabarti D., Wang Y., Wang C. [et al.]. Epidemic thresholds in real networks // ACM Transactions on Information and System Security (TISSEC). 2008. Vol. 10 (4). P. 1–26. DOI: 10.1145/1284680.1284681.

4. Giaever G., Chu A. M., Ni L. [et al.]. Functional profiling of the Saccharomyces cerevisiae genome // Nature. 2002. Vol. 418. P. 387–391. DOI: 10.1038/nature00935.

5. Said M. R., Begley T. J., Oppenheim A. V. [et al.]. Global network analysis of phenotypic effects: protein networks and toxicity modulation in Saccharomyces cerevisiae // Proceedings of the National Academy of Sciences of the United States of America. 2004. Vol. 101 (52). P. 18006–18011. DOI: 10.1073pnas.0405996101.

6. Ciriello G., Gatza M. L., Beck A. H. [et al.]. Comprehensive molecular portraits of invasive lobular breast cancer // Cell. Vol. 163, Issue 2. P. 506–519. DOI: 10.1016/j.cell.2015.09.033.

7. Chatr-Aryamontri A., Oughtred R., Boucher L. [et al.]. The BioGRID interaction database: 2017 update // Nucleic Acids Research. 2017. Vol. 45, Issue D1. P. D369–D379. DOI:10.1093/nar/gkw1102.

8. Sun S., Ling L., Zhang N. [et al]. Topological structure analysis of the protein-protein interaction network in budding yeast // Nucleic Acids Research. 2003. Vol. 31, no. 9. P. 2443–2450.

9. Le S., Riva A., Tran D. A high-performance pipeline for genome-wide network reconstruction from gene expression data // Proceedings of NetBio SIG, ISMB 2016, Orlando, FL, USA. URL: http://public.ndexbio.org/#/network/3bb11a95-dace-11e7-adc1-0ac135e8bacf (дата обращения: 28.09.2018).

10. Huang J. K., Carlin D. E., Yu M. K. [et al.]. Systematic evaluation of molecular networks for discovery of disease genes // Cell Systems. 2018. Vol. 4 (6). P. 484–495.e5. DOI: 10.1016/j.cels.2018.03.001.

11. Pratt D., Chen J., Welker D. [et al.]. NDEx, the network data exchange // Cell systems. 2015. Vol. 1, Issue 4. P. 302–305. DOI: 10.1016/j.cels.2015.10.001.

12. Sun S., Ling L., Zhang N. [et al.]. Topological structure analysis of the protein-protein interaction network in budding yeast // Nucleic Acids Research. 2003. Vol. 31, no. 9. P. 2443–2450. DOI: 10.1093/nar/gkg340.

13. Задорожный В. Н. Растущие сети: динамика распределения степеней связности смежных узлов // Омский научный вестник. 2016. № 2 (146). С. 81–86.

14. Задорожный В. Н., Юдин Е. Б. Калибровка случайных графов предпочтительного связывания по распределениям степеней вершин и ребер // Омский научный вестник. 2017. № 1 (151). С. 114—118.

15. Posternak V., Cole M. D. Strategically targeting MYC in cancer // F1000Research. 2016. Vol. 5. DOI:10.12688/f1000research.7879.1.

16. Kashtan N., Itzkovitz S., Milo R. [et al.]. Topological generalizations of network motifs // Physical Review E 70. 2004. P. 031909-1—031909-12. DOI: 10.1103/PhysRevE.70.031909.

17. Milo R., Shen-Orr S. S., Itzkovitz S. [et al.]. Network motifs: Simple building blocks of complex networks // Science. 2002. Vol. 298 (5594). P. 824—827. DOI: 10.1126/science.298.5594.824.

18. Schreiber F., Schwöbbermeyer H. Frequency concepts and pattern detection for the analysis of motifs in networks // Transactions on Computational Systems Biology III. Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 2005. Vol. 3737. P. 89—104. DOI: 10.1007/11599128_7.

19. Wernicke S., Rasche F. FANMOD: a tool for fast network motif detection. PDF // Bioinformatics. 2006. Vol. 22 (9). P. 1152—1153. DOI: 10.1093/bioinformatics/btl038.

20. Chen J., Hsu W., Li Lee M. [et al.]. NeMoFinder: dissecting genome-wide protein-protein interactions with meso-scale network motifs // The 12th ACM SIGKDD international conference on Knowledge discovery and data mining. Philadelphia, Pennsylvania, USA. 2006. P. 106—115. DOI: 10.1145/1150402.1150418.

21. Kashani Z. R., Ahrabian H., Elahi E. [et al.]. Kavosh: a new algorithm for finding net-work motifs // BMC Bioinformatics. 2009. Vol. 10 (318). DOI: 10.1186/1471-2105-10-318.

22. Meira Luis A., Vinicius A., Maximo R. [et al.]. Ace-Motif: accelerated network motif detection // IEEE/ACM Computational Biology and Bioinformatics. 2014. Vol. 11, no. 5. P. 853—862. DOI: 10.1109/TCBB.2014.2321150.

23. Yudin E. B., Zadorozhnyi V. N. Statistical approach to calculation of number of network motifs // 2015 International Siberian Conference on Control and Communications (SIBCON). 2015. 7147296. DOI: 10.1109/SIBCON.2015.7147296.

24. Yudin E. B., Yudina M. N. Calculation of number of motifs on three nodes using random sampling of frames in networks with directed links // 2017 Siberian Symposium on Data Science and Engineering (SSDSE). 2017. 8071957. P. 23—26. DOI: 10.1109/ssdse.2017.8071957.

25. Zadorozhnyi V. N., Yudin E. B., Yudina M. N. Graphs with complex stochastic increments // 2017 IEEE Dynamics of Systems, Mechanisms and Machines (Dynamics), 14—16 November, Omsk, 2017. P. 500—508.

26. Юдина М. Н. Информационный образовательный ресурс локального доступа «Библиотека классов MotifsLib»: свидетельство о регистрации электронного ресурса № 23487 от 19.02.2018 // Хроники Объединенного фонда электронных ресурсов «Наука и образование». 2018. № 2 (105). С. 22. DOI: 10.12731/ofernio.2018.23487.

27. Юдина М. Н. Информационный образовательный ресурс локального доступа «Библиотека классов SocAndBioNetworksAnalysis»: свидетельство о регистрации электронного ресурса в № 23770 от 15.09.2018 // Хроники Объединенного фонда электронных ресурсов «Наука и образование». 2018. № 9 (112). С. 25. DOI: 10.12731/ofernio.2018.23770.

28. Юдина М. Н. Информационный образовательный ресурс локального доступа «Библиотека классов YMN_GraphGenerators»: электронного ресурс № 23769 от 15.09.2018 // Хроники Объединенного фонда электронных ресурсов «Наука и образование». 2018. № 9 (112). С. 25. DOI: 10.12731/ofernio.2018.23769.

ЮДИНА Мария Николаевна, аспирантка кафедры «Автоматизированные системы обработки информации и управления».
SPIN-код: 3000-9556
ORCID: 0000-0002-9648-6409
AuthorID (SCOPUS): 57195502392
ResearcherID: R-4589-2016
Адрес для переписки: mg-and-all@mail.ru

Для цитирования

Юдина М. Н. Комплекс программных библиотек для анализа молекулярных сетей клетки // Омский научный вестник. 2018. № 6 (162). С. 265—270. DOI: 10.25206/1813-8225-2018-162-265-270.

Статья поступила в редакцию 05.10.2018 г.