Вавиловский журнал генетики и селекции. 2022;26(8):826-829
DOI 10.18699/VJGB-22-100

Перевод на английский язык https://vavilov.elpub.ru/jour

Свойства малого мира научных организаций определяют динамику публикационной активности в области миРНК

А.Б. Фирсов[1] ✉, И.И. Титов[2, 3]

[1] Институт систем информатики им. А.П. Ершова Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[3] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
✉ artyomfirsov@mail.ru

Аннотация. Многие научные статьи стали доступны в цифровом виде, что позволяет запрашивать данные статей и, в частности, автоматически собирать метаданные, включая данные об аффилиации. Это, в свою очередь, можно использовать для количественных оценок научной области, например для идентификации организаций и анализа графа соавторства этих организаций для извлечения базовой структуры науки. В настоящей работе рассмотрена область исследования микроРНК, а именно граф соавторства организаций и анализ его эволюции. Чтобы решить проблему вариативности написания названия организаций, был предложен алгоритм сортировки логических векторов признаков k-mer/n-gram. В нем используется тот факт, что содержание аффилиации довольно консистентно для одной и той же организации. Для учета ошибок написания и других артефактов названия организации в поле метаданных аффилиации наш подход преобразует упоминание организации внутри аффилиации в K-Mer (n-gram) булевый вектор присутствия. Далее векторы всех аффилиаций из набора данных лексикографически сортируются, образуя группы упоминаний организаций. Таким подходом был кластеризован набор данных аффилиаций в области исследования микроРНК и определены названия уникальных организаций, что позволило построить граф соавторства на уровне научных организаций. С помощью этого графа показано, что рост области исследования микроРНК контролируется архитектурой малого мира сети научных организаций и испытывает степенной рост с показателем степени 2.64 ± 0.23 для числа организаций в соответствии с диаметром сети, предлагая модель роста новых научных направлений. Скорость публикации первой статьи по микроРНК у организации при ее взаимодействии с другой организацией, уже публиковавшейся в этой области, аппроксимируется как 0.184 ± 0.002 год$^{-1}$.

Ключевые слова: k-mer; n-gram; миРНК; электронная библиотека; соавторство организаций; малый мир.

Для цитирования: Фирсов А.Б., Титов И.И. Свойства малого мира научных организаций определяют динамику публикационной активности в области миРНК. *Вавиловский журнал генетики и селекции*. 2022;26(8):826-829. DOI 10.18699/VJGB-22-100

Small world of the miRNA science drives its publication dynamics

A.B. Firsov[1] ✉, I.I. Titov[2, 3]

[1] A.P. Ershov Institute of Informatics Systems of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[3] Novosibirsk State University, Novosibirsk, Russia
✉ artyomfirsov@mail.ru

Abstract. Many scientific articles became available in the digital form which allows for querying articles data, and specifically the automated metadata gathering, which includes the affiliation data. This in turn can be used in the quantitative characterization of the scientific field, such as organizations identification, and analysis of the co-authorship graph of those organizations to extract the underlying structure of science. In our work, we focus on the miRNA science field, building the organization co-authorship network to provide the higher-level analysis of scientific community evolution rather than analyzing author-level characteristics. To tackle the problem of the institution name writing variability, we proposed the k-mer/n-gram boolean feature vector sorting algorithm, KOFER in short. This approach utilizes the fact that the contents of the affiliation are rather consistent for the same organization, and to account for writing errors and other organization name variations within the affiliation metadata field, it converts the organization mention within the affiliation to the K-Mer (n-gram) Boolean presence vector. Those vectors for all affiliations in the dataset are further lexicographically sorted, forming groups of organization mentions. With that approach, we clustered the miRNA field affiliation dataset and extracted unique

organization names, which allowed us to build the co-authorship graph on the organization level. Using this graph, we show that the growth of the miRNA field is governed by the small-world architecture of the scientific institution network and experiences power-law growth with exponent 2.64 ± 0.23 for organization number, in accordance with network diameter, proposing the growth model for emerging scientific fields. The first miRNA publication rate of an organization interacting with already publishing organization is estimated as 0.184 ± 0.002 year^{-1}.
Key words: k-mer; n-gram; miRNA; digital library; organization co-authorship; small world.

For citation: Firsov A.B., Titov I.I. Small world of the miRNA science drives its publication dynamics. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2022;26(8):826-829. DOI 10.18699/VJGB-22-100

Введение

Научные структуры стимулируют продуктивность научной работы, обеспечивая исследователей материально-техническими условиями и научной средой. Один из факторов эффективности научной работы – взаимодействие ученых в виде обмена идеями или совместной работы, которое проявляется в виде соавторства научных публикаций. Анализ соавторства исследовательских институтов, а не характеристик на уровне авторов дает возможность обеспечить более высокий уровень анализа эволюции научного сообщества, в частности организацию из «невидимых колледжей» или развитие международного сотрудничества в глобальном масштабе (Leydesdorff et al., 2013). Подобные исследования направлены на поиск причин конкуренции и сотрудничества в конкретных областях (Wagner, Leydesdorff, 2005), а также на выявление закономерностей международной публикационной активности (Ribeiro et al., 2018). В целом для понимания структуры научного сообщества и процесса распространения знаний в области науки анализ должен проводиться как на уровне авторов, так и на уровне организаций.

Граф имеет свойство малого мира, если $L \propto \log(N)$, где L – среднее кратчайшее расстояние графа, N – количество вершин графа. Другими словами, любые две вершины достижимы из другой посредством малого количества переходов через другие вершины, но при этом вероятность того, что они смежные, мала.

Такой тип сетей встречается во многих явлениях: например, распространение инфекции (Liu et al., 2015), нейронные связи (Muldoon et al., 2016) и др. Отдельный интерес представляет анализ эффекта малого мира в распространении знаний (Shi, Guan, 2016), и поэтому в нашей работе проверяется, является ли граф взаимодействия организаций в области исследования микроРНК малым миром.

Поскольку в малом мире вершины достижимы друг до друга за малое количество переходов, такие процессы, как распространение инфекции или знания, должны происходить иначе, чем в обычном графе.

Для определения того, что граф является малым миром, в нескольких работах были предложены различные критерии (Watts, Strogatz, 1998; Newman et al., 2000). В нашей работе мы выбрали категориальный критерий для выявления присутствия малого мира в сети организаций микроРНК, следуя (Humphries, Gurney, 2008), где авторы вводят меру «малости мира»:

$$S = \frac{CC_{\mathrm{G}}}{CC_{\mathrm{rand}}} \Big/ \frac{L_{\mathrm{G}}}{L_{\mathrm{rand}}} .$$

В уравнении выше CC_{G} – коэффициент кластеризации графа G; L_{G} – средняя длина кратчайшего пути графа G; CC_{rand} и L_{rand} – параметры случайного графа со случайно равномерным размещением ребер с тем же количеством узлов и ребер, что и граф G.

Процесс распространения знаний можно интерпретировать как процесс «заражения идеями», при котором через промежуточного хозяина (научные публикации) организации могут вдохновиться какой-либо областью исследований и сами начать публиковать статьи. Такой процесс можно моделировать с помощью модели Susceptible, Infectious, Recovered (SIR) (Goffman, Newill, 1964), в рамках которой составляется система дифференциальных уравнений, моделирующих динамику заражения и выздоровления субъектов. В простейшем случае однородной среды решением этих уравнений на малых временах является экспоненциальный рост числа зараженных субъектов.

В работе (Vazquez, 2006) автор моделирует рост заболеваемости с использованием SIR для задач, где графы передачи известны и обладают свойством малого мира (Muldoon et al., 2016). Автор адаптирует модель распространения SIR к представлению исходного графа в виде остовного дерева (AST) и получает точную нормализованную частоту заболеваемости для AST, $\rho(t)$, которая аппроксимирует эту частоту для исходного графа. Таким образом, учитывая, что граф обладает свойством малого мира, существует точное решение нормализованной частоты заражения для AST, которое является аппроксимацией для исходного графа:

$$\rho(t) = \lambda \frac{(\lambda t)^{D-1}}{(D-1)!} e^{-(\lambda+\mu)t} \left[1 + O\left(\frac{t_0}{t}\right)\right],$$

где λ и μ – соответственно скорости заражения и выздоровления в рамках SIR модели; D – среднее кратчайшее расстояние графа; t_0 – время перехода между режимами. Для этого граф, помимо характеристики малого мира, должен удовлетворять одному из условий для γ (показатель степенного закона распределения степеней вершин) и υ (коэффициент корреляции Пирсона степени между парами соединенных узлов) (Vazquez, 2006):

$$\gamma > 3, \quad \upsilon > 0,$$
$$2 \le \gamma \le 3, \quad \upsilon > -1, \quad 3 - \gamma + \upsilon > 0.$$

Материалы и методы

Для сбора набора данных аффилиаций области исследования микроРНК использовалась цифровая библиотека PubMed. Из аффилиаций были выделены упоминания организаций. При этом использовался подход, основан-

ный на ключевых словах, для понимания, какая часть аффилиации содержит какую информацию об упоминании организации (название организации, страна, город и т. д.):

Пример разбиения аффилиации на упоминания организаций с определением страны для статьи с PubMed ID 19996210

(1) Authors' Affiliations: Cancer Genetics, Kolling Institute of Medical Research; Department of Endocrinology; Department of Anatomical Pathology, Royal North Shore Hospital, St. Leonards, New South Wales, Australia; Department of Surgery, Bankstown Hospital, Bankstown, New South Wales, Australia; South Western Sydney Clinical School, University of New South Wales; Endocrine Surgical Unit, University of Sydney; Department of Surgery, Liverpool Hospital, Sydney, New South Wales, Australia; Endocrine Surgical Unit, University of California Los Angeles; and Division of Hematology and Oncology, Department of Medicine, University of California Los Angeles School of Medicine, Los Angeles, California.	1. kolling institute of medical research, Australia 2. royal north shore hospital, Australia 3. bankstown hospital, Australia 4. university of new south wales, Australia 5. university of sydney, Australia 6. liverpool hospital, Australia 7. university of california los angeles, UNKNOWN 8. university of california los angeles, school of medicine, UNKNOWN

Затем для всех упоминаний был построен словарь уникальных K-Mer (n-gram), где K = 2, и для каждого упоминания сформирован булевый вектор присутствия определенного K-Mer в этом упоминании. Далее эти векторы упоминаний были отсортированы лексикографически, чтобы получить список векторов, в котором схожие упоминания сгруппированы по построению. После чего для каждой соседней пары упоминаний было посчитано расстояние по метрике Dice. Если оно превышало заданный порог, это было свидетельством того, что упоминания относятся к различным кластерам, что дает нам группировку упоминаний (см. таблицу).

Сгруппированные упоминания содержат ссылки на одну и ту же организацию, поэтому на следующем шаге мы можем построить граф соавторства организаций, определяя, какие организации опубликовали одну и ту же статью вместе.

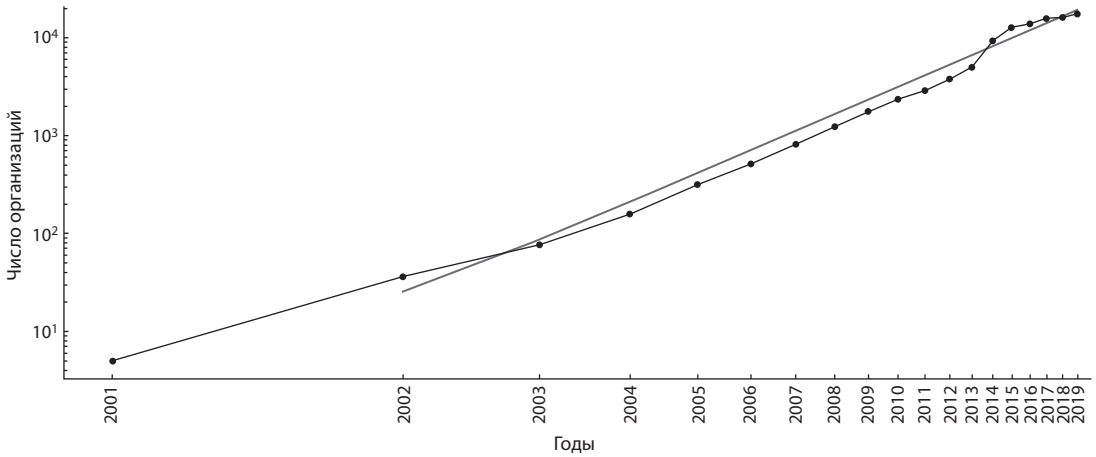
Результаты

Анализ структурных характеристик графа научных организаций в области исследования микроРНК показывает, что этот граф удовлетворяет критериям малого мира (Muldoon et al., 2016) с показателем степени степенного распределения $\gamma = 2.01$ и коэффициентом ассортативности связности вершин графа $\nu = -0.03$. Поэтому для числа научных организаций, имеющих публикации в данной области, можно ожидать степенной рост согласно модели (Vazquez, 2006). Из (Vazquez, 2006) следует, что начальный рост числа вершин имеет степенную зависимость с показателем степени $D - 1$, где D – средняя длина кратчайшего пути в графе. Для графа научных организаций области исследования микроРНК $D = 3.46$, а аппроксимированный степенной параметр $D - 1 = 2.64 \pm 0.23$ (см. рисунок), что

Пример идентификации организаций

#	Упоминание	2-Mer булевый вектор	Метрика Dice
1	institute	**1111111**00000000	0.2
2	insitute	**11111001**00001000	0.429
3	institue	**1111011**0000010000	**0.834**
4	center	0000100**011100100**	0.4
5	centre	0000000**011100011**	

Примечание. Значение порога – 0.8, K = 2. Расстояние между элементами 3, 4 превышает значение порога, что приводит к разбиению элементов на кластеры. Примеры 2-Mer – in, ns, st, ti, it, tu, …



Ежегодное количество организаций, опубликовавших статью в области исследования микроРНК в зависимости от времени в двойных логарифмических координатах.

для показателя степени дает отклонение около 7 % от предсказанного моделью.

При аппроксимации скорости «заражения информацией» получена величина скорости $\lambda = 0.184 \pm 0.002$ год$^{-1}$, которая характеризует скорость публикации первой статьи по микроРНК у организации при ее взаимодействии с другой организацией, уже публиковавшейся в этой области.

Анализ подграфа российских научных институтов в области исследования микроРНК показывает, что активность российских организаций уступает средней активности организаций (среднее количество публикаций на одну организацию в России составляет 0.92 против среднего по области значения 21.5). При этом российское сообщество оказалось более плотным: коэффициент кластеризации подграфа российских организаций превышает средний по области: 0.708 в России и 0.361 в среднем. Самый активный партнер России в международном сотрудничестве – США, с 50 совместными публикациями. Однако американо-российское сотрудничество нестабильно и децентрализовано, а лидерами по активности сотрудничества с российскими организациями являются Немецкий центр изучения рака, Харбинский медицинский университет и Каролинский институт (по 6 совместных публикаций).

Обсуждение

Понимание факторов продуктивности исследовательских организаций и динамики их публикационной активности важно для управления наукой. Помимо алгоритмов автоматической идентификации организаций, активно развиваются такие проекты, как ror.org, направленные на идентификацию научных институтов за счет присвоения им уникальных идентификаторов (подобно orcid.org для авторов). Данные проекты упрощают поиск организаций, но требуют принятия использования таких проектов авторами публикаций, так как для возможности полной идентификации каждой организации необходимо указывать идентификатор ror.org для каждой аффилиации из публикации, что на текущий момент не может быть гарантировано. Поэтому в ближайшее время алгоритмы автоматической идентификации организаций останутся актуальными.

В работе данные представлены по 2019 г. К настоящему времени структура графа могла измениться. Кроме того, информация о статьях в библиотеке PubMed может обновляться ретроспективно. Тем не менее данные по публикациям на 23.01.2022 показывают, что картина эволюции области микроРНК принципиально не изменилась (данные не приведены). Новая геополитическая реальность неизбежно скажется на структуре взаимодействия и соавторства в научных областях. Однако в связи с запаздыванием по времени видимых результатов изменение в научном сотрудничестве проявится в базах данных не ранее 2024 г.

Заключение

Одной из моделей развития новых областей знания является модель «интеллектуальной эпидемии», в которой новые идеи случайно распространяются среди исследователей, заражая все большее и большее их число (Goffman, Newill, 1964). Закон распространения может определяться структурой среды. В нашей работе показано, что граф взаимодействия организаций области исследования микроРНК является малым миром, и вследствие этого публикационная активность области демонстрирует степенной рост согласно модели (Vazquez, 2006). Более медленный по сравнению с экспоненциальным, степенной рост возникает из-за «самоизбегания» путей распространения в компактных сетях малого мира: при «заражении информацией» очередного узла малого мира высока вероятность того, что этот узел уже был «заражен» альтернативным путем. Граф соавторства для нашего анализа был построен с использованием алгоритма кластеризации упоминаний организаций на основе сортировки булевых векторов признаков K-Mer.

Список литературы / References

Goffman W., Newill V.A. Generalization of epidemic theory. An application to the transmission of ideas. *Nature*. 1964;204(4955):225-228. DOI 10.1038/204225a0.

Humphries M.D., Gurney K. Network ‘small-world-ness’: a quantitative method for determining canonical network equivalence. *PLoS One.* 2008;3(4):e0002051. DOI 10.1371/journal.pone.0002051.

Leydesdorff L., Wagner C., Park H., Adams J. International collaboration in science: the global map and the network. *Prof. Inf.* 2013;22(1):1-18. DOI 10.3145/epi.2013.ene.12.

Liu M., Li D., Qin P., Liu C., Wang H., Wang F. Epidemics in interconnected small-world networks. *PLoS One.* 2015;10(3):e0120701. DOI 10.1371/journal.pone.0120701.

Muldoon S., Bridgeford E., Bassett D. Small-world propensity and weighted brain networks. *Sci. Rep.* 2016;6:22057. DOI 10.1038/srep22057.

Newman M.E.J., Moore C., Watts D.J. Mean-field solution of the small-world network model. *Phys. Rev. Lett.* 2000;84(14):3201-3204. DOI 10.1103/PhysRevLett.84.3201.

Ribeiro L., Rapini M., Silva L., Albuquerque E.M. Growth patterns of the network of international collaboration in science. *Scientometrics.* 2018;114:159-179. DOI 10.1007/s11192-017-2573-x.

Shi Y., Guan J. Small-world network effects on innovation: evidences from nanotechnology patenting. *J. Nanopart. Res*. 2016;18:329. DOI 10.1007/s11051-016-3637-1.

Vazquez A. Spreading dynamics on small-world networks with connectivity fluctuations and correlations. *Phys. Rev. E. Stat. Nonlin. Soft Matter Phys*. 2006;74:056101. DOI 10.1103/PhysRevE.74.056101.

Wagner C., Leydesdorff L. Network structure, self-organization and the growth of international collaboration in science. *Res. Policy*. 2005;34(10):1608-1618. DOI 10.1016/j.respol.2005.08.002.

Watts D.J., Strogatz S.H. Collective dynamics of ‘small-world’ networks. *Nature.* 1998;393(6684):440-442. DOI 10.1038/30918.

ORCID ID

A. Firsov orcid.org/0000-0002-7681-1032
I.I. Titov orcid.org/0000-0002-2691-3292

Благодарности. Работа ИТ была поддержана в рамках государственного бюджетного проекта РФ ФВНР-2022-0020.

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию 07.09.2022. После доработки 10.11.2022. Принята к публикации 10.11.2022.