

Интеллектуальная эпидемия в науке о миРНК

Фирсов А.^{1*}, Титов И.^{2,3}

¹ Институт систем информатики им. А.П. Ершова СО РАН, Новосибирск, Россия

² Институт цитологии и генетики СО РАН, Новосибирск, Россия

³ Новосибирский государственный университет, Новосибирск, Россия

* artemijfirsov@mail.ru

Key words: *k*-мер; *n*-gram; dbscan; идентификация; микроРНК; timsort; kofer; электронная библиотека; соавторство организаций; малый мир; степенной рост

Мотивация и цель: Цифровая форма научных статей позволяет автоматически запрашивать их метаданные, которые могут быть использованы для оценки статистики научной области. В нашей работе мы сосредоточились на области науки о миРНК, для которой мы построили сеть соавторства организаций, чтобы обеспечить анализ на более высоком уровне [1] эволюции научного сообщества по сравнению с характеристиками авторов. В данной работе мы показываем, что развитие области науки о миРНК описывается эпидемиологической моделью [2], которая расширяет распространение модели SIR на трансмиссионный график. Ранее мы показали, что рост области миРНК подчиняется архитектуре малого мира [3]. В этой работе мы предоставляем анализ степенного роста публикаций в этой области, а также показываем сравнение на уровне стран, выделяя США и Китай как лидеров.

Распространение знания как эпидемия: Болезни распространяются, когда инфекционный материал встречает восприимчивое население, состоящее из Susceptible, Infectious, or Recovered. Воздействие может привести к иммунитету или инфицированию с латентным периодом до появления симптомов. Эпидемии возникают, когда наблюдается всплеск заболеваний. Гофман [4] расширил это, обсуждая эпидемии идей, такие как религиозные или научные концепции. В отличие от желательных эпидемий идей, вспышки заболеваний нежелательны. В научных или лаконичных населенных пунктах уязвимые члены могут заразиться концепцией, в то время как иммунные индивиды могут отвергнуть ее или распространять ее ненамеренно. Когда значительная часть населения принимает концепцию, происходит эпидемия знаний.



Рис. 1. Переход от восприимчивого к инфицированному и к выздоровевшему состоянию в модели SIR

Такая концепция была применена к популяции с известным трансмиссионным графом в [2], что обсуждается в следующем разделе.

Динамика сетей малого мира: Распространение знаний можно сравнить с эпидемией, где организации уязвимы для «заражения» определенными научными областями и впоследствии публикуются в них. Эти динамики часто представляются с использованием моделей Susceptible, Infectious, or Recovered (SIR) [6], которые демонстрируют экспоненциальный рост инфекций на начальных стадиях. Тем не менее эти модели не учитывают основной граф передачи заболевания. В [2] автор моделирует рост заболеваемости для проблем, где известны графы передачи и имеется свойство малого мира [7], более подробное объяснение было дано в [3], и полученная модель для «зараженных» организаций:

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

где λ/μ – параметры экспоненциации заражения/выздоровления из модели SIR, D – средний кратчайший путь графа, t_0 время переключения между степенным ростом и замедлением экспоненты.

Методы и алгоритмы: Наш подход включал использование цифровой библиотеки PubMed для компиляции набора данных по исследованиям миРНК. Аффиляции, упомянутые в этих статьях, были сгруппированы с помощью лексикографической сортировки предварительно нормализованных и разделенных векторов признаков К-мер булева типа. Эти сгруппированные аффиляции содержат ссылки на уникальные организации, что облегчает построение графа соавторства для выявления организаций, сотрудничавших при публикациях.

Результаты: Как было показано ранее [3], граф миРНК обладает свойством малого мира, и с $D = 3.46$, и аппроксимированным параметром экспоненциации $D - 1 \approx 2.6$, мы видим отклонение на 5.5 % от прогноза модели.

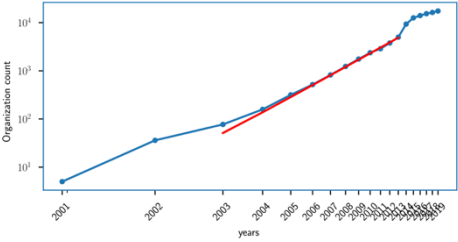


Рис. 2. Ежегодно учитываются организации, опубликовавшие статью в течение определенного года

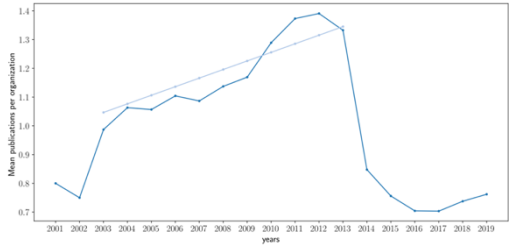


Рис. 3. Среднее количество публикаций на организацию в год

В данной работе мы показываем, что среднее количество публикаций на организацию следует линейному росту, который может быть аппроксимирован и использован для сравнения точной динамики публикаций с оцененной с использованием модели степенного закона, а также динамики среднего количества публикаций на организацию:

$$a_{org} * x_{org}^{b_{org}} * (mx + l) = a * t^b,$$

где $mx + l$ – среднее количество публикаций на организацию в год, b_{org} – параметр экспоненциации степенного роста организаций, b – параметр экспоненциации степенного роста публикаций, a_{org}, a – константы степенной аппроксимации роста организаций и публикаций соответственно.

Расчеты показывают, что точное количество публикаций имеет вид $a * t^b = 4.16e - 04 * t^{2.85}$, тогда как количество публикаций, оцененное на основе данных об организациях и среднем количестве публикаций, составляет $a * t^b = 2.78e - 04 * t^{2.82}$, что приводит к отклонению около 1 %. Это в свою очередь показывает, что микросвойства области миРНК согласуются с макросвойствами, и структура взаимодействия организаций в определенной степени определяет активность публикаций.

С учетом публикаций, также имеющих модель степенного закона, мы сравнили ведущие страны и показали, что США имели быстрый старт публикаций в этой области, достигая примерно 1000 публикаций в год в 2006 г (5 лет после первого года публикации – 2001 г.). В то время как Китай имеет более высокий общий рост, что привело к тому, что Китай стал новым лидером в 2013 г. Кроме того, большинство лидеров испытало снижение активности публикаций в последние 3–4 года.

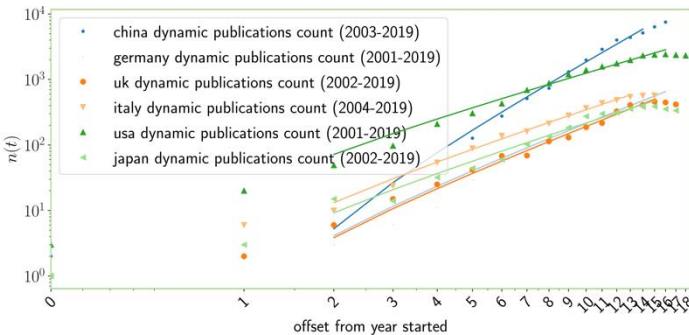


Рис. 4. Ежегодные публикации США, Китая, Германии, Италии, Великобритании, Японии в науке о миРНК со степенными приближениями. 0 – это год, когда страна опубликовала первую работу. Эти годы можно увидеть в легенде рисунка. Первые две и последние две точки были исключены из приближения. Данные каждой страны нормализованы к максимальному числу кумулятивных публикаций этой конкретной страны

Заключение: В данной работе мы показали, что граф взаимодействия научных учреждений, связанных с миРНК, следует эпидемической модели [2] и организован через эффект малого мира, а также демонстрирует степенной рост, что тоже проявляется на уровне публикаций. Основываясь на аппроксимации степенным законом, мы показали, что лидерами этой области являются США и Китай, причем последний является текущим лидером.

miRNA science as an intellectual epidemic

Firsov A.^{1*}, Titov I.^{2, 3}

¹ A.P. Ershov Institute of Informatics Systems, SB RAS, Novosibirsk, Russia

² Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia

Key words: k -mer; n -gram; dbscan; identification; miRNA; timsort; kofer; digital library; organization co-authorship; small world; power growth; epidemic

Motivation and Aim: Digital form of scientific articles allows for automatic querying of their metadata, which can be used in estimating the statistics of the scientific field. In our work, we focus on the miRNA science field, for which we built the organization co-authorship network to provide higher-level analysis [1] of scientific community evolution compared to the author-level characteristics. In this work we show that the development of the miRNA science field is described by the epidemic model [2], that extends SIR model spreading onto transmission graph. Previously we showed that the growth of the miRNA field is governed by the small-world architecture [3], and in this work we provide the analysis of the power-law growth of publications within the area, and show the country-level comparison, outlining USA and China as the leaders.

Epidemic Knowledge Spreading: Diseases spread when infective material meets a susceptible population, consisting of infectives, susceptibles, and removals. Exposure can lead to immunity or infection, with a latency period before symptoms appear. Epidemics arise when there's a spike in cases. Goffman and Newill [4] expanded on this, discussing idea epidemics like religious or scientific concepts. Unlike desirable idea epidemics, disease outbreaks are undesirable. In scientific or lay populations, vulnerable members can catch a concept, while immune individuals may reject it or spread it unknowingly. When a significant portion of the population adopts a concept, a knowledge epidemic occurs. Such concept was further developed and applied to the transmission graphs in [2], which is discussed in the next section.



Fig. 2. Transition from Susceptible to Infectious to Recovered states in the SIR model

Dynamics of the Small-world Networks: The dissemination of knowledge can be likened to a contagion, wherein organizations are vulnerable to being “infected” by certain research fields and subsequently producing papers. These dynamics are often represented using Susceptible, Infectious, or Recovered (SIR) models [6], which demonstrate exponential growth in infections during initial stages. Nevertheless, these models do not account for the underlying graph of disease transmission.

In [2], the author models the incidence growth for problems, where the transmission graphs are known and have the small-world property [7], more detailed explanation was given in [3], and the resulting model for “infected” organizations is:

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

where λ/μ – exponent parameters of infection/removal events of the SIR model, D – average shortest path length of the graph, t_0 – crossover time between power-law growth and exponent deceleration.

Methods and Algorithms: Our approach involved utilizing the PubMed digital library to compile a dataset on miRNA research. Affiliations mentioned in these papers were clarified by employing lexicographical sorting on pre-normalized and split K-Mer Boolean feature vectors. These clustered affiliations encompass references to identical organizations, facilitating the construction of a co-authorship graph to identify organizations that collaborated on publications.

Results: As shown previously [3], the miRNA graph has the small-world property, and with $D = 3.46$, and the approximated exponentiation parameter $D - 1 \approx 2.6$, it shows 5.5 % deviation from the model's predicted.

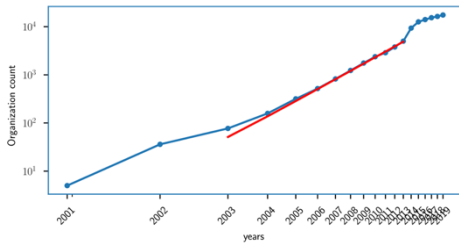


Fig. 2. Yearly organizations count, which published a paper within a certain year

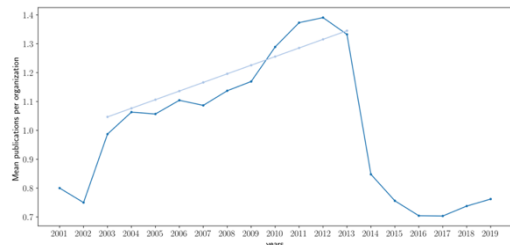


Fig. 3. Average publications per organization per year

In this work, we show that the average publications per organization follow a linear growth pattern, which can be approximated and used to compare the exact publication dynamics with estimated using the power-law model, and the dynamics of the average publications per organization:

$$a_{org} * x_{org}^{b_{org}} * (mx + l) = a * t^b,$$

where $mx + l$ – average publications per organization per year, b_{org} – organizations count exponentiation parameter, b – publications count exponentiation parameter, a_{org}, a – constants of power approximation of organizations and publications respectively.

The calculations show that the exact publications count has $a * t^b = 4.16e - 04 * t^{2.85}$, whereas publications count estimated based on the organizations and average publications data gives $a * t^b = 2.78e - 04 * t^{2.82}$, resulting in the deviation to be around 1 %, which shows that micro properties of the miRNA field are consistent with the macro properties, and the small-world structure of organization interaction defines the publication activity.

And with publications shown to also support the power-law model, we compared the leader countries using this model, showing that the USA had the rapid start of publication in this field, reaching approximately 1,000 publications per year in 2006 (5 years after the first year of publication – 2001). Whereas China has higher overall growth, which led to China becoming the new leader in 2013. Additionally, most of the leaders experienced the decline of publication activity in the late 3–4 years.

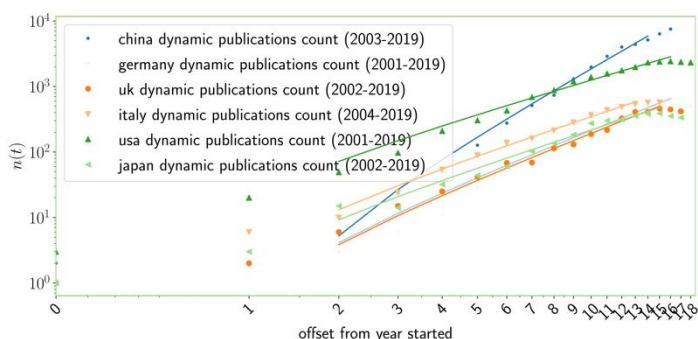


Fig. 4. Yearly publications of USA, China, Germany, Italy, UK, Japan in the miRNA science per year with power approximations. 0 is the year the country had the first publication. These years can be seen in the legend section of the figure. First two and last two points were excluded from the approximation. The data of each country was normalized to the maximal number of cumulative publications of this specific country

Conclusion: In this work we showed that the miRNA’s scientific institution interaction graph follows the epidemic model [2] and is organized via the small-world effect and shows the power-law growth, which also manifests itself on the publications level. Based on the power-law approximation, we showed that the leaders of the field are USA, and China, with the latter being the current leader.

Список литературы/References

1. Leydesdorff L., Wagner C., Park H., Adams J. International collaboration in science: The global map and the network. *El Profesional de la Informacion*. 2013;22(1):87-94. doi 10.3145/epi.2013.ene.12
2. Vazquez A. Spreading dynamics on small-world networks with connectivity fluctuations and correlations. *Physical Rev*. 2006;74:056101
3. Firsov A., Titov I. Small world effect of the miRNA science field drives its growth. In: Bioinformatics of Genome Regulation and Structure/Systems Biology (BGRS/SB-2022). Novosibirsk: ICG SB RAS, 2022;1090-1091. doi 10.18699/BGRS/SB-2022-000
4. Goffman W. Mathematical approach to the spread of scientific ideas – the history of mast cell research. *Nature*. 1966;212(5061):449-452. doi 10.1038/212449a0
5. Ribeiro L., Rapini M., Silva L., Motta Albuquerque E. Growth patterns of the network of international collaboration in science. *Scientometrics*. 2018;114:159-179
6. Beckley R., Weatherspoon C., Alexander M. et al. Modeling epidemics with differential equations. 2013. [<https://www.tnstate.edu/mathematics/mathreu/filesreu/GroupProjectSIR.pdf>]
7. Humphries M.D., Gurney K. Network ‘small-world-ness’: A quantitative method for determining canonical network equivalence. *PLoS One*. 2008;3(4):e0002051