Progress and Knowledge Transfer from Science to Technology in the Research Frontier of CRISPR Based on the LDA Model

Yushuang Lyu¹, Muqi Yin², Fangjie Xi¹, Xiaojun Hu^{1†}

¹Medical Information Center, and Department of Neurology of Affiliated Hospital 2, Zhejiang University School of Medicine, Hangzhou 310058, China

²Institute of Cyber-Systems and Control, Zhejiang University School of Control Science and Engineering, Hangzhou 310007, China

Citation: Lyu, Y.S., Yin, M.Q., Xi, F.J., & Hu, X.J. (2022). Progress and knowledge transfer from science to technology in the research frontier of CRISPR based on the LDA model. *Journal of Data and Information Science*, 7(1), 1–19. <https://doi.org/10.2478/jdis-2022-0004>

Received: Oct. 20, 2021

Revised: Dec. 30, 2021

Accepted: Dec. 31, 2021

Abstract

Purpose: This study explores the underlying research topics regarding CRISPR based on the LDA model and figures out trends in knowledge transfer from science to technology in this area over the latest 10 years.

Design/methodology/approach: We collected publications on CRISPR between 2011 and 2020 from the Web of Science, and traced all the patents citing them from lens.org. 15,904 articles and 18,985 patents in total are downloaded and analyzed. The LDA model was applied to identify underlying research topics in related research. In addition, some indicators were introduced to measure the knowledge transfer from research topics of scientific publications to IPC-4 classes of patents.

Findings: The emerging research topics on CRISPR were identified and their evolution over time displayed. Furthermore, a big picture of knowledge transition from research topics to technological classes of patents was presented. We found that for all topics on CRISPR, the average first transition year, the ratio of articles cited by patents, the NPR transition rate are respectively 1.08, 15.57%, and 1.19, extremely shorter and more intensive than those of general fields. Moreover, the transition patterns are different among research topics.

Research limitations: Our research is limited to publications retrieved from the Web of Science and their citing patents indexed in lens.org. A limitation inherent with LDA analysis is in the manual interpretation and labeling of “topics”.

Practical implications: Our study provides good references for policy-makers on allocating scientific resources and regulating financial budgets to face challenges related to the transformative technology of CRISPR.

[†] Correspondence author: Xiaojun Hu (E-mail: xjhu@zju.edu.cn).

Research Paper

Originality/value: The LDA model here is applied to topic identification in the area of transformative researches for the first time, as exemplified on CRISPR. Additionally, the dataset of all citing patents in this area helps to provide a full picture to detect the knowledge transition between S&T.

Keywords CRISPR; LDA model; Knowledge transfer; Transformative technology

1 Introduction

The prokaryote-derived CRISPR (clustered regularly interspaced short palindromic repeats) is a genetic engineering technique in molecular biology by which the genomes of living organisms, including humans, may be modified (Ledford, 2015). It enables targeted genetic modifications in cultured cells, as well as whole animals and plants with extremely high precision, cheaply, and with ease (Doudna, 2020; Kim & Kim, 2014; Zhu, Li, & Gao, 2020). Especially, the CRISPR/Cas9 genetic scissors discovered in 2012 by Emmanuelle Charpentier and Jennifer A. Doudna who won the 2020 Nobel Prize in Chemistry have not only revolutionized basic science but also resulted in innovative crops and will lead to ground-breaking new medical treatments (<http://www.nobelprize.org/>). The diversity, modularity, and efficacy of CRISPR-Cas systems are driving a biotechnological revolution and the field of CRISPR-based biotechnology is developing at a rapid pace (Doudna, 2020; Doudna & Charpentier, 2014; Pickar-Oliver & Gersbach, 2019).

The impact of recent development of CRISPR on science and biotechnology is immense (Doudna & Gersbach, 2015). Academic research in this area has made great progress especially since its breakthroughs in 2012 (Knott & Doudna, 2018). A few investigations aimed at assessing the development and status of this field have been conducted, for example, Qin, Wang, and Ye (2019) conducted a research to reveal the main research hotspots of CRISPR/Cas9 by constructing and analyzing core keyword co-occurrence networks. Zhou et al. (2021) studied the evolution of academic research hotspots in CRISPR by analysing the usage of key phrases. However, most of them are limited to general bibliometric analysis based on a small amount of selected data. Today, an increased interest has been shown in using topic modeling methods to explore the structure of a certain domain (Figuerola, Marco, & Pinto, 2017; Han, 2020; Lamba & Madhusudhan, 2019; Sugimoto et al., 2011). The method presents a broader and more comprehensive perspective based on a large number of publications.

Science research is also regarded as a driving force for technology innovation and economic growth (Gittelman & Kogut, 2003; Li, Azoulay, & Sampat, 2017; Lo, 2010). New basic knowledge may lead to a knowledge transfer to technological fields resulting from interactions of citations, and, in this way, promotes innovations

in industry and progress in society (Hu & Rousseau, 2018). Besides, this incredible technique over our heritable information, CRISPR, also brings unknown risks to human beings. Thus, it becomes very important to track and monitor the technological applications of CRISPR, which can also help to ensure responsible innovations by scientists and use it in an ethical and safe way while driving the technology forward (Baltimore et al., 2015; Gurwitz, 2014).

In this article, we present a new perspective on observing and understanding the research topics and their development in CRISPR based on the LDA model and try to illustrate the knowledge transfer trends from research topics to technological fields based on paper-patent citations.

2 Methodology

2.1 LDA model

Topic modeling methods have been frequently employed to detect the intellectual structure of a scientific domain based on a large number of publications as the corpus (Han, 2020). They consist of statistical techniques and aim to describe the topics in the documents and the relations between those topics and their evolution over time (Blei, 2012). Latent Dirichlet Allocation (LDA), proposed by Blei, Ng, and Jordan (2003), is one of the best known and most widely used models. It is a classical unsupervised-learning algorithm most widely applied to mining the potential topics of a large number of texts. This model is based on a three-layer Bayesian probabilistic model containing probabilistic relations among the structure of words, topics, and text. Given a text set, it represents each text as its corresponding topic set and each topic as a particular polynomial distribution of words (Zhou et al., 2019). LDA has natural advantages in identifying emerging research topics and zooming into special areas of research (Mendes et al., 2019; Suominen & Toivanen, 2016).

There are multiple implementations of LDA. In this article, we applied the Gensim[®], a python library, to perform LDA. The LDA in Gensim is with an online variational Bayes (VB) algorithm developed by Hoffman, Bach, and Blei (2010).

2.2 Determining the number of topics

It is important to identify the “correct” number of topics in an LDA model (Arun et al., 2010). Too many topics dilute the meaning of each topic and too few do not allow publications to be distinguished from each other (Lamba & Madhusudhan, 2019; Kushkowsk et al., 2020). There are several proposals to optimize the number of topics utilizing computational methods, such as perplexity (Blei, Ng, & Jordan,

[®] <https://pypi.org/project/gensim/>

2003), coherence (Roder et al., 2015), and so on. Some researchers choose the number of topics by human judgment (Figuerola, Marco, & Pinto, 2017; Newman & Block, 2006). Here the appropriate number of topics for our dataset is determined by comparing the inter-topic similarity with LDAvis.

LDAvis (Sievert & Shirley, 2014), a web-based interactive visualization tool for qualitative assessment of topic models, enables an intuitive, yet profound, inspection of topic-term relationships in an LDA model (Miyata et al., 2020). The tool represents the inter-topic similarity as the distance between the topics, which is by calculating inter-topic Jensen-Shannon divergence (JS divergence) values. If the JS divergence value is larger, the topic similarity is smaller, i.e., the distance between the two topics is further. Thus, a good model would have as few overlapping topics as possible. (Sievert & Shirley, 2014; Chehal et al., 2021). We performed pyLDAvis (the python package for LDAvis) for the model visualization in this paper.

2.3 Knowledge transition indicators

Based on all publications related to CRISPR and their citing patents, we introduce some indicators to detect knowledge transfer from science to technology in this area.

(1) The first transition year

The first transition year (FT) (Hu & Rousseau, 2018; van Raan, 2017b) represents the number of years it takes for an article to receive its first citation by a patent and it would be zero if the article is cited by a patent in its publication year. It reflects the speed of technological impact of an article (van Raan, 2017b).

(2) NPR transition rate

NPR transition rate is originally defined by Hu and Rousseau (2018), we introduce it and make a slight change in the context of this contribution, that is, the number of citing patents divided by the number of articles on the corresponding topic. It reflects to what extent articles on some topic contribute to the technological world.

(3) Relative strength of knowledge transition

Hu and Rousseau (2018) developed relative strength of knowledge transition (SKT) to analyze the characteristics from scientific fields to technological classes. Here, we use the notion “research topics” to replace “scientific fields”, if $P(i)$ is one patent class of the citing patents and $A(j)$ is one research topic in the article set then the SKT between $P(i)$ and $A(j)$ is defined as:

$$SKT_{ij} = \frac{C_{ij}}{T}$$

Where C_{ij} indicates the number of citations from patents in class $P(i)$ to articles in topic $A(j)$. The sum of all weighted citations between P and article set (any article

under study), denoted as T . Note that if a patent belongs to two (or more) patent classes, its citation is counted twice or more, i.e., whole counting. Likewise, the SKT of a patent class $P(i)$ is:

$$SKT_i = \frac{C_i}{T}$$

Where C_i represents the number of citations from patents in $P(i)$ to any article under study. We, moreover, define the SKT of $A(j)$:

$$SKT_j = \frac{C_j}{T}$$

Here C_j denotes the total number of citations received by articles that belong to topic $A(j)$, coming from all patent classes.

3 Data source and collection

Publications covering the genome editing toolbox based on CRISPR systems such as Cas9, Cas12, Cas13, base editors (BEs), and prime editors (PEs) were drawn from Clarivate Analytics' Web of Science (WoS) Core Collection in August 2021. Specifically, we used the following search string.

TS= ("CRISPR*" OR "clustered regularly interspaced short palindromic repeats*" OR "base editor*" OR "base editing" OR "prime editing" OR ((Cas9 OR Cas10 OR Cas11 OR Cas12 OR Cpf1 OR Cas13 OR Cas14 OR CasX OR gRNA OR sgRNA) AND ("gene editing" OR "genome engineering" OR "genome editing" OR "genome editor" or "genome binding"))))

First, we performed data cleanup and removed irrelevant data. Final data were limited to the article type, and included publications with a title, an abstract, keywords and the publication year. This totaled 15,904 records between 2011 and 2020. The title, keywords and abstract text were saved in a Microsoft Excel file for the LDA model, and the publication year was used to incorporate the LDA results for topic changes analysis.

Next, we searched by article title on lens.org to trace their patent citations one by one. For articles that have been cited by patents, we downloaded their citing patents and assigned them a unique identifier to link them to their citing patents. In this way we obtained for each article the number of patent citations, the application year, and the corresponding International Patent Classification codes (IPC codes) for each patent. Note that, we only counted the number of different families to avoid double counting. Totally, we obtained 18,985 citing patent families of all articles. Then we developed a Python script to get the first four digits of the IPC code, which is the IPC-4 code dividing the patent into a subclass level (WIPO, 2019). After that, we

Research Paper

constructed a matrix of relationships between scientific topics and IPC-4 codes (based on their citation links). Here, only different IPC-4 codes were counted. Moreover, Python, Microsoft Excel, Gephi, and Tableau were applied to data visualization.

4 Results

4.1 The explosive growth of publications and keywords on CRISPR

As shown in Fig. 1, this transformative technology has triggered an explosive growth in publications and keywords co-occurrence networks over the last decade, especially since 2012. These years witness an explosive growth in the number of publications from 68 in 2011 to 4,295 in 2020. It can also be seen that the keywords co-occurrence network has been booming from 593 keywords in 2011–2012 to 16,074 in 2019–2020.

Figure 1. Growth of publication numbers (a) and keywords co-occurrence networks (b) regarding CRISPR in 2011–2020.

4.2 Identification of research topics

We have evaluated models with different numbers of topics (1 to 50 topics) and refer to the LDAvis for visualization of each model as well. When the number of topics is 10, the topics are distributed in all quadrants with relatively few connections, which means that the models are relatively independent and have little mutual

similarity, see Fig. 2. Therefore, the model with 10 topics is selected as the optimal model that fits our corpus the best.

As the model visualization shown in Fig. 2, It depicts a global topic view of the model on the left, and the term bar charts reveal the top 30 most relevant terms on the right, among which topics 1 and 4 are relatively close, followed by topic 8 and 10.

Figure 2. The layout of topics on CRISPR by the method of LDAvis (an example with topic 1 selected).

Table 1 summarizes the LDA results for the CRISPR articles. It may be considered as the major research interests over the last decade. In addition, the top 20 words are further listed, ranked by the probability value in descending order.

4.3 Topic evolution analysis

To determine the topic of an article, we chose the topic with greater presence, following the LDA results in that article. We can then represent the evolution of these topics, based on the number of publications produced each year in the last decade, see Fig. 3.

One trend to notice is that the growth rate of publication numbers between 2011 and 2014 is slow. But beginning with 2014, the total number of articles on CRISPR topics increases steadily with the larger amount of publications. There is no doubt that it is caused by the advent of the transformative technique, the CRISPR/Cas9 genetic scissors, in 2012. In general, the number of articles on all topics has been

Research Paper

Table 1. 10 topics and corresponding words on CRISPR.

No	Topic	Top 20 most correlated words
#1	Gene mutation based on CRISPR/Cas9	mice; gene; development; mutations; model; function; mutant; cas9; protein; loss; disease; crispr; mutation; expression; knockout; type; mutations; using; results; mouse
#2	Genome editing based on CRISPR system	crispr; genome; based; gene; can; genetic; editing; cas9; system; target; using; high; engineering; new; method; rna; single; technology; screening; tools
#3	Targeted therapies	beta; expression; alpha; protein; induced; cells; signaling; receptor; stress; pathway; activation; response; increased; mediated; role; level; mir; regulation; factor; dependent
#4	Gene expression	expression; protein; transcription; rna; gene; binding; cell; proteins; virus; dna; replication; regulation; transcriptional; promotor; infection; auxin; activation; complex; factor; human
#5	Human therapeutic	cell; cancer; tumor; expression; proliferation; growth; knockout; crispr; resistance; inhibition; survival; cas9; apoptosis; lines; human; treatment; migration; patients; therapeutic; lung
#6	Biotechnology with CRISPR/Cas9	cas9; editing; crispr; dna; genome; target; gene; system; efficiency; mutations; mediated; base; repair; recombination; guide; using; efficient; homologous; single; double
#7	Genetic engineering human cells	cell; gene; stem; human; cas9; crispr; delivery; expression; gfp; pluripotent; vivo; using; editing; protien; derived; system; retinal; mouse; knock; therapy
#8	CRISPR in agriculture	crispr; rice; resistance; plant; detection; clustered; short; palindromic; regularly; interspaced; repeats; analysis; results; associated; study; pneumoniae; arabidopsis; pcr; toxin; clinical
#9	Genome analysis on strains	genes; strains; strain; production; acid; genome; study; analysis; resistance; identified; biosynthesis; growth; species; genomic; isolated; metabolic; clusters; two; showed; genomes
#10	Mechanism of CRISPR/Cas	crispr; cas; systems; dna; phage; bacterial; system; rna; proteins; type; sequence; cleavage; host; immunity; pam; complex; anti; immune; structure; plasmids

Figure 3. The number of publications in each topic from 2011 to 2020 (per year).

on the rise in the past ten years, which also provides evidence about the large influence of CRISPR on scientific research. In addition, the different growth rates reflect the development of topics. Among them, topics 1 (Gene mutation based on CRISPR/Cas9) and 2 (Genome editing based on CRISPR system) have reached the top 2 and became the most important topics since 2016.

Figure 4. The percentage of publications in each topic from 2011 to 2020 (per year).

Fig. 4 shows the percentage of the total output of publications that come from each topic, further illustrating the evolution of topics in CRISPR over time. Specifically, topic 10 (Mechanism of CRISPR/Cas) has contributed a large share (about 50%) since the beginning and has been able to maintain its position until 2014. It is reasonable that the emergence of CRISPR as a transformative technology first brings the in-depth studies of its mechanism, which have also laid the foundation for follow-up researches. Then new topics emerge over time with topic 6 and topic 7 appearing in 2013, as well as topic 1 and 5 in 2014, i.e., the rapid development of CRISPR broadens its research applications into “Gene mutation based on CRISPR/Cas9”, “Human therapeutic”, “Biotechnology with CRISPR/Cas9” and “Genetic engineering human cells”. Gradually, the proportion of some topics decreases, such as topics 6 (Biotechnology with CRISPR/Cas9) and 10 (Mechanism of CRISPR/Cas), while the relative contribution of some topics expands, such as topic 1 (Gene mutation based on CRISPR/Cas9) and 5 (Human therapeutic). As time goes by, the proportions of each topic remain stable. Besides, similar to the results in Fig. 3, the dominant position of topic 10 at the beginning has been replaced by topics 1 and 2 since 2016.

4.4 Knowledge transfer trends from science to technology

Science research is proved to be the driving force behind technology development and is important for promoting technological innovations (Fukuzawa & Ida, 2016; Lo, 2010; McMillan, Narin & Deeds, 2000). Patent citations to research articles offer a way to identify contributions of scientific knowledge to technological development (Tijssen, Buter, & van Leeuwen, 2000). Thus, we aim to reveal the knowledge transfer trends between S&T in CRISPR based on direct citation analysis.

As a whole, 2,477 of 15,904 articles (a share of 15.57%) have been cited by 18,985 patents over the last decade in this area. The NPR transition rate and Avg. FT (the average first transition year of articles) in each topic are shown in table 2. The smallest 3 values in Avg. FT and the largest 3 in other columns are bolded.

Table 2. Detailed data on paper-patent citations in CRISPR.

	Publications	NPRs	% of NPRs	Citing patents	NPR transition rate	Avg. FT (years)
topic 1	3,019	53	1.76%	512	0.17	0.87
topic 2	3,118	439	14.08%	4,443	1.42	1.12
topic 3	1,392	19	1.36%	149	0.11	1.11
topic 4	1,425	159	11.16%	1,220	0.86	1.18
topic 5	1,751	27	1.54%	290	0.17	0.63
topic 6	1,687	600	35.57%	5,926	3.51	1.07
topic 7	1,028	121	11.77%	1,252	1.22	1.07
topic 8	473	26	5.50%	204	0.43	0.88
topic 9	855	26	3.04%	107	0.13	1
topic 10	1,156	1,007	87.11%	4,882	4.22	1.09
total	15,904	2,477	15.57%	18,985	1.19	1.08

(1) Very fast knowledge transition speeds in CRISPR

All topics have an average first transition year (FT) within 1.18 years in table 2, which is quite atypical compared to the result from related studies in which the average citation age from science research to technological development is 9.8 years in the area of genetic editing (Lo, 2010). This result indicates that the knowledge transfer speed from science to technology is very fast in CRISPR. In other words, articles on CRISPR, are quickly cited by patents, especially for articles concerning topic 5 (Human therapeutic) with the smallest value of 0.63.

(2) High but unbalanced NPR transition rates

As table 2 illustrates, about 15.57% of the publications in CRISPR are cited by patents, while van Raan discovered that only a small minority of publications covered by the WoS are cited by patents, about 3%–4% (van Raan, 2017a). This means that publications on CRISPR are relatively high technology-relevant.

Moreover, the overall NPR transition rate is 1.19, which even exceeds the value of applications-oriented research which was 0.392 in biotechnology found by Hu and Rousseau (2018). This reflects the tremendous technological influence of articles regarding CRISPR. Specifically, the NPR transition rates vary by research topic. The values of topic 10 (Mechanism of CRISPR/Cas) and 6 (Biotechnology with CRISPR/Cas9) are far above the others with respective values are 4.22 and 3.51, while the values of topics 3, 9, 1, and 5 are below average. Their respective values are only 0.11, 0.13, 0.17 and 0.17. According to Tijssen (2010), the roles of articles in knowledge transfer are dependent on the degree of application orientation in them. Topics with high NPR transition rates tend to play major roles in knowledge transfer and they are somewhat more likely to drive technological innovation in CRISPR.

(3) Knowledge transfer strength from research topics to technological classes

From the technological perspective, 87 out of 632 IPC-4 codes in total are transferred from scientific articles in CRISPR. These codes cover 6 technological sections from A to H, except section D (Textiles; Paper) and E (Fixed Constructions) (WIPO, 2019). In terms of each topic, topic 2 (Genome editing based on CRISPR system) covers the largest scope with 60 IPC-4 codes. This indicates that articles concerning “Genome editing based on CRISPR system” have the widest technological impacts. Overall, the top 10 IPC-4 codes ranked by their SKT values contribute 95%. We refer to “Appendix” for their detailed explanations (WIPO, 2019).

As displayed in Fig. 5, the knowledge flows between S&T are illustrated with a Sankey diagram in which the width of inks is proportional to the SKT value. It not only shows how knowledge moves between topics and top 10 IPC-4 codes but also helps to understand the quantitative dependency between research topics and technological classes in CRISPR. The top 3 total outflows are topics 6, 10, and 2 with SKT values 0.31, 0.25, and 0.23, which means that most technological applications depend highly on the topics “Genome editing based on CRISPR system”, “Biotechnology with CRISPR/Cas9” and “Mechanism of CRISPR/Cas”.

In addition, table 3 shows the corresponding SKT value of Fig. 5 and the top 3 largest values of each topic (in bold). The results suggest that the C12N and A61K domains are the top 2 inflows from each topic, particularly, C12N accounts for the main part of inflows with a high SKT value of 0.46.

Figure 5. Knowledge transfer from topics of articles to IPC-4 codes of patents.

Table 3. Relative strength of knowledge transition (SKT) between research topics and technological classes.

	C12N	A61K	C07K	C12Q	A61P	A01K	C12P	G01N	C07H	A01H	Others	Sum
topic 1	0.0122	0.0067	0.0034	0.0016	0.0029	0.0022	0.0001	0.0010	0.0005	0.0004	0.0014	0.03
topic 2	0.1070	0.0322	0.0212	0.0191	0.0103	0.0066	0.0057	0.0038	0.0037	0.0031	0.0168	0.23
topic 3	0.0023	0.0026	0.0007	0.0003	0.0020	0.0001	0.0001	0.0002	0.0001	0.0000	0.0015	0.01
topic 4	0.0278	0.0112	0.0064	0.0061	0.0039	0.0010	0.0012	0.0019	0.0019	0.0008	0.0031	0.07
topic 5	0.0038	0.0049	0.0020	0.0012	0.0032	0.0003	0.0000	0.0010	0.0001	0.0001	0.0016	0.02
topic 6	0.1493	0.0423	0.0286	0.0225	0.0128	0.0135	0.0058	0.0051	0.0047	0.0069	0.0143	0.31
topic 7	0.0291	0.0144	0.0083	0.0038	0.0042	0.0026	0.0007	0.0015	0.0007	0.0002	0.0039	0.07
topic 8	0.0040	0.0015	0.0005	0.0019	0.0006	0.0004	0.0001	0.0003	0.0001	0.0006	0.0007	0.01
topic 9	0.0018	0.0011	0.0003	0.0003	0.0004	0.0001	0.0005	0.0001	0.0001	0.0001	0.0017	0.01
topic 10	0.1213	0.0361	0.0275	0.0191	0.0105	0.0060	0.0042	0.0032	0.0054	0.0047	0.0140	0.25
Sum	0.46	0.15	0.1	0.08	0.05	0.03	0.02	0.02	0.02	0.02	0.05	1

4 Discussion & conclusion

In this contribution, we applied the LDA model for topic detection in the transformative area, CRISPR, and demonstrated the development of topics over time. Results show that the dominant topics have gradually evolved from “Mechanism of CRISPR/Cas” to “Gene mutation based on CRISPR/Cas9” and “Genome editing based on CRISPR system”. Moreover, by tracking all patent citations of articles, we

discovered that the publications in this area are highly technology relevant and affect the technological world at an extremely rapid pace. Technological influence varies with research topics, among which the topics “Mechanism of CRISPR/Cas”, “Biotechnology with CRISPR/Cas9” and “Genome editing based on CRISPR system” form top 3. Finally, as shown in Fig. 5, we figured out a big picture of knowledge transfer between S&T on CRISPR, where the C12N and A61K domains are the two most important inflows from each topic.

As a transformative technique, CRISPR attracted considerable attention from worldwide scientists and international organizations. Although many experts have described the development and applications of CRISPR from the perspective of qualitative analysis, such as recapitulating genetic mutations in animals or cellular models (Hsu, Lander, & Zhang, 2014), programmable RNA targeting and viral gene disruption (Doudna & Charpentier, 2014), DNA changes in stem cells and treating human diseases (Baltimore et al., 2015), yet they just focused on limited aspects of CRISPR in a traditional mode of review (Hsu, Lander, & Zhang, 2014; Doudna & Charpentier; 2014; Baltimore et al., 2015). In contrast, our contribution provided more detailed and objective results including the research topics in CRISPR and their development based on a large number of publications in the past decade. Hence, a full picture of the new technology was taken up, which also helps scholars grasp the current research trends and seize the opportunity of scientific research on CRISPR.

The pathway of knowledge transfer is a main bridge for understanding the interaction between science and technology. It is seen as an essential source of innovation and a mechanism for the dissemination of research results (Campbell et al., 2020; Wang & Ye, 2021). By measuring knowledge transfer through direct paper-patent citations in the frontier of CRISPR, we found that scientific publications and their impact on the technological field vary by research topic. The results not only provide meaningful clues for knowledge management in this domain, but also benefit policy-makers in formulating scientific and technological policies, as well as allocating scientific resources. For example, the SKT value of “Biotechnology with CRISPR/Cas9” is 0.31, while SKT values of “CRISPR in agriculture”, “Targeted therapies” and “Genome analysis on strains” are only 0.01. The largest knowledge flow (SKT value of 0.1493) is from the topic “CRISPR/Cas9 biotechnology” to technological field “C12N”. The findings above are valuable for the decision-making of research foundations, e.g. funders can decide whether to support the technology-relevant topics or the basic-oriented topics according to the scientific or economic development goals.

The rapid development and broad application prospects of CRISPR also come with a great responsibility to use it ethically and safely (Doudna & Gersbach, 2015).

Research Paper

This simple and widely available technology can now be used to perform genome modification in eggs, sperm or embryos, thereby changing the genetic makeup of human germline, human beings are facing unknown risks in science, medicine, and ethics (Doudna, 2020; Baltimore et al., 2015). Thus, the application of this technology, especially topics on “Human therapeutic” and “Genetic engineering human cells” in the above results, must be rigorously regulated. Our research provides a good reference for regulators in tracking the technological application of CRISPR to drive the technology forward while ensuring responsible use.

Many studies have confirmed that LDA can effectively cluster meaningful and interpretable topics from a large number of documents (Blei & Lafferty, 2007; Suominen & Toivanen, 2016; Yau et al., 2014). In today’s world, various extensions of LDA have been proposed to further detect topic changes over time, such as the Topic Over Time model (TOT) and the Dynamic Topic Model (DTM) (Blei & Lafferty, 2006; Shan & Li, 2010; Wang & McCallum, 2006). As a transformative technology, the publications on CRISPR has been explosively growing in the latest decade (as shown in Fig. 1). Hence in this study, we applied the time post-discretized analysis to detect the topic changes, on the basis of the number of articles produced each year (Griffiths & Steyvers, 2004; Shan & Li, 2010), that is, running the LDA model based on the entire text set and then incorporating the publication year of texts into the LDA results. This method has been serving as a reliable means to gain insight into the dynamics of science (Figuerola, Marco, & Pinto, 2017; Griffiths & Steyvers, 2004; Shan & Li, 2010; Jiang et al., 2021). The result of this work may be beneficial for scientists, intelligence analysts and policy-makers to meet the challenges related to this disruptive technology.

Our study has limitations. We only focused on publications related to CRISPR retrieved from the Web of Science and their citing patents indexed in lens.org. Besides, a limitation inherent with LDA analysis is in the manual interpretation and labeling of “topics”. Some topics are fairly straightforward to label by interpreting the word distribution and examining the most representative articles of them in detail, while others are proved more difficult to ascertain the content or relationship that connected the words and articles. We hope that further research will provide additional insights built on our work.

Acknowledgments

This work was supported by the National Natural Science Foundation of China, Grant numbers: 71974167 and 71573225. The authors would like to thank the anonymous reviewers for their constructive comments.

Author contributions

Yushuang Lyu (21918174@zju.edu.cn): data collection, data processing, and analysis, programming, the design of methodology, writing the manuscript. Muqi Yin (21932043@zju.edu.cn): data processing and programming. Fangjie Xi (22018070@zju.edu.cn): data collection and data processing. Xiaojun Hu (xjhu@zju.edu.cn): initiated the idea, research question proposal, the design of methodology, writing the manuscript.

References

- Arun, R., Suresh, V., Madhavan, C.E.V., & Murty, M.N. (2010). On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations. In M.J. Zaki, J.X. Yu, B. Ravindran, & V. Pudi (Eds.), *Advances in Knowledge Discovery and Data Mining, Pt I, Proceedings* (Vol. 6118, pp. 391–402). Berlin: Springer-Verlag Berlin.
- Baltimore, D., Berg, P., Botchan, M., Carroll, D., Charo, R.A., Church, G., ... Yamamoto, K.R. (2015). A prudent path forward for genomic engineering and germline gene modification. *Science*, 348(6230), 36–38. doi:10.1126/science.aab1028
- Blei, D.M., Ng, A.Y., & Jordan, M.I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(4/5), 993–1022
- Blei, D.M. (2012). Probabilistic topic models. *Communications of the Acm*, 55(4), 77–84. doi:10.1145/2133806.2133826
- Blei, D.M., & Lafferty, J.D. (2006). Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning* (pp. 113–120).
- Blei, D.M., & Lafferty, J.D. (2007). A correlated topic model of science. *Annals of Applied Statistics*, 1(1), 17–35. doi:10.1214/07-aos114
- Campbell, A., Cavalade, C., Haunold, C., Karanikic, P., & Piccaluga, A. (2020). Knowledge Transfer Metrics. Towards a European-wide set of harmonised indicators. *EUR*, 30218
- Doudna, J.A. (2020). The promise and challenge of therapeutic genome editing. *Nature*, 578(7794), 229–236. doi:10.1038/s41586-020-1978-5
- Doudna, J.A., & Charpentier, E. (2014). The new frontier of genome engineering with CRISPR-Cas9. *Science*, 346(6213), 1077–+. doi:10.1126/science.1258096
- Doudna, J.A., & Gersbach, C.A. (2015). Genome editing: The end of the beginning. *Genome Biology*, 16. doi:10.1186/s13059-015-0860-5
- Figuerola, C.G., Marco, F.J.G., & Pinto, M. (2017). Mapping the evolution of library and information science (1978–2014) using topic modeling on LISA. *Scientometrics*, 112(3), 1507–1535. doi:10.1007/s11192-017-2432-9
- Fukuzawa, N., & Ida, T. (2016). Science linkages between scientific articles and patents for leading scientists in the life and medical sciences field: The case of Japan. *Scientometrics*, 106(2), 629–644. doi:10.1007/s11192-015-1795-z
- Gittelman, M., & Kogut, B. (2003). Does good science lead to valuable knowledge? Biotechnology firms and the evolutionary logic of citation patterns. *Management Science*, 49(4), 366–382. doi:10.1287/mnsc.49.4.366.14420
- Griffiths, T.L., & Steyvers, M. (2004). Finding scientific topics. In *Proceedings of the National Academy of Sciences of the United States of America*, 101, 5228–5235. doi:10.1073/pnas.0307752101

Research Paper

- Gupta, P., & Gulati, P. (2021). Implementation and comparison of topic modeling techniques based on user reviews in e-commerce recommendations. *Journal of Ambient Intelligence and Humanized Computing*, 12(5), 5055–5070. doi:10.1007/s12652-020-01956-6
- Gurwitz, D. (2014). Gene drives raise dual-use concerns. *Science*, 345(6200), 1010–1010. doi:10.1126/science.345.6200.1010-b
- Han, X.Y. (2020). Evolution of research topics in LIS between 1996 and 2019: An analysis based on latent Dirichlet Allocation topic model. *Scientometrics*, 125(3), 2561–2595. doi:10.1007/s11192-020-03721-0
- Hsu, P.D., Lander, E.S., & Zhang, F. (2014). Development and applications of CRISPR-Cas9 for genome engineering. *Cell*, 157(6), 1262–1278. doi:10.1016/j.cell.2014.05.010
- Hoffman, M., Bach, F., & Blei, D. (2010). Online learning for latent dirichlet allocation. *Advances in Neural Information Processing Systems*, 23, 856–864.
- Hu, X., & Rousseau, R. (2018). A new approach to explore the knowledge transition path in the evolution of science & technology: From the biology of restriction enzymes to their application in biotechnology. *Journal of Informetrics*, 12(3), 842–857. doi:10.1016/j.joi.2018.07.004
- Jiang, T., Liu, X.P., Zhang, C., Yin, C.A.H., & Liu, H.Z. (2021). Overview of trends in global single cell research based on bibliometric analysis and LDA model (2009–2019). *Journal of Data and Information Science*, 6(2), 163–178. doi:10.2478/jdis-2021-0008
- Kim, H., & Kim, J.S. (2014). A guide to genome engineering with programmable nucleases. *Nature Reviews Genetics*, 15(5), 321–334. doi:10.1038/nrg3686
- Knott, G.J., & Doudna, J.A. (2018). CRISPR-Cas guides the future of genetic engineering. *Science*, 361(6405), 866–869. doi:10.1126/science.aat5011
- Kushkowski, J.D., Shrader, C.B., Anderson, M.H., & White, R.E. (2020). Information flows and topic modeling in corporate governance. *Journal of Documentation*, 76(6), 1313–1339. doi:10.1108/jd-10-2019-0207
- Lamba, M., & Madhusudhan, M. (2019). Mapping of topics in DESIDOC Journal of Library and Information Technology, India: A study. *Scientometrics*, 120(2), 477–505. doi:10.1007/s11192-019-03137-5
- Ledford, H. (2015). CRISPR, the disruptor. *Nature*, 522(7554), 20–24. doi:10.1038/522020a
- Li, D., Azoulay, P., & Sampat, B.N. (2017). The applied value of public investments in biomedical research. *Science*, 356(6333), 78–81. doi:10.1126/science.aal0010
- Lo, S.C.S. (2010). Scientific linkage of science research and technology development: A case of genetic engineering research. *Scientometrics*, 82(1), 109–120. doi:10.1007/s11192-009-0036-8
- McMillan, G.S., Narin, F., & Deeds, D.L. (2000). An analysis of the critical role of public science in innovation: The case of biotechnology. *Research Policy*, 29(1), 1–8. doi:10.1016/s0048-7333(99)00030-x
- Mendes, F.M.L., Castor, K., Monteiro, R., Mota, F.B., & Rocha, L.F.M. (2019). Mapping the lab-on-a-chip patent landscape through bibliometric techniques. *World Patent Information*, 58. doi:10.1016/j.wpi.2019.101904
- Miyata, Y., Ishita, E., Yang, F., Yamamoto, M., Iwase, A., & Kurata, K. (2020). Knowledge structure transition in library and information science: Topic modeling and visualization. *Scientometrics*, 125(1), 665–687. doi:10.1007/s11192-020-03657-5

- Newman, D.J., & Block, S. (2006). Probabilistic topic decomposition of an eighteenth-century American newspaper. *Journal of the American Society for Information Science and Technology*, 57(6), 753–767. doi:10.1002/asi.20342
- Pickar-Oliver, A., & Gersbach, C.A. (2019). The next generation of CRISPR-Cas technologies and applications. *Nature Reviews Molecular Cell Biology*, 20(8), 490–507. doi:10.1038/s41580-019-0131-5
- Qin, J.H., Wang, J.J., & Ye, F.Y. (2019). A metric approach to hot topics in biomedicine via keyword co-occurrence. *Journal of Data and Information Science*, 4(4), 13–25. doi:10.2478/jdis-2019-0018
- Roder, M., Both, A., Hinneburg, A., & Assoc Comp, M. (2015). Exploring the space of topic coherence measures. New York: Assoc Computing Machinery.
- Shan B. & Li F. (2010). A survey of topic evolution based on LDA (in Chinese). *Journal of Chinese Information Processing*, 24(06), 43-49+68
- Sievert, C., & Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces* (pp. 63–70).
- Sugimoto, C.R., Li, D.F., Russell, T.G., Finlay, S.C., & Ding, Y. (2011). The shifting sands of disciplinary development: Analyzing north american library and information science dissertations using Latent Dirichlet Allocation. *Journal of the American Society for Information Science and Technology*, 62(1), 185–204. doi:10.1002/asi.21435
- Suominen, A., & Toivanen, H. (2016). Map of science with topic modeling: Comparison of unsupervised learning and human-assigned subject classification. *Journal of the Association for Information Science and Technology*, 67(10), 2464–2476. doi:10.1002/asi.23596
- Tijssen, R.J.W. (2010). Discarding the ‘basic science/applied science’ dichotomy: A knowledge utilization triangle classification system of research journals. *Journal of the American Society for Information Science and Technology*, 61(9), 1842–1852. doi:10.1002/asi.21366
- Tijssen, R.J.W., Buter, R.K., & van Leeuwen, T.N. (2000). Technological relevance of science: An assessment of citation linkages between patents and research papers. *Scientometrics*, 47(2), 389–412. Retrieved from <Go to ISI>://WOS:000089449100014
- Wang, J.J., & Ye, F.Y. (2021). Probing into the interactions between papers and patents of new CRISPR/CAS9 technology: A citation comparison. *Journal of Informetrics*, 15(4), 12. doi:10.1016/j.joi.2021.101189
- Wang, X., & McCallum, A. (2006). Topics over time: A non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 424–433).
- WIPO. (2019). International patent classification (version 2019). Available at <http://www.wipo.int/classifications/ipc>
- van Raan, A.F.J. (2017a). Patent citations analysis and its value in research evaluation: A review and a new approach to map technology-relevant research. *Journal of Data and Information Science*, 2(1), 13–50. doi:10.1515/jdis-2017-0002
- van Raan, A.F.J. (2017b). Sleeping beauties cited in patents: Is there also a dormitory of inventions? *Scientometrics*, 110(3), 1123–1156. doi:10.1007/s11192-016-2215-8

Research Paper

- Yau, C.K., Porter, A., Newman, N., & Suominen, A. (2014). Clustering scientific documents with topic modeling. *Scientometrics*, 100(3), 767–786. doi:10.1007/s11192-014-1321-8
- Zhou, H.C., Zheng, D.J., Li, Y.M., & Shen, J.W. (2019). User-opinion mining for mobile library apps in China: Exploring user improvement needs. *Library Hi Tech*, 37(3), 325–337. doi:10.1108/lht-05-2018-0066
- Zhou, W.Y., Yuan, Y.J., Zhang, Y.Q., & Chen, D. (2021). A decade of CRISPR gene editing in China and beyond: A scientometric landscape. *Crispr Journal*, 4(3), 313–320. doi:10.1089/crispr.2020.0148
- Zhu, H.C., Li, C., & Gao, C.X. (2020). Applications of CRISPR-Cas in agriculture and plant biotechnology. *Nature Reviews Molecular Cell Biology*, 21(11), 661–677. doi:10.1038/s41580-020-00288-9

This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivs License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Appendix

Table 4. Detailed explanations of the top 10 IPC-4 classes.

Rank	IPC-4 class	Detailed meaning
1	C12N	Microorganisms or Enzymes; Compositions Thereof; Propagating, Preserving, or Maintaining Microorganisms; Mutation or Genetic Engineering; Culture Media;
2	A61K	Preparations for Medical, Dental, or Toilet Purposes;
3	C07K	Peptides;
4	C12Q	Measuring or Testing Processes Involving Enzymes, Nucleic Acids or Microorganisms; Compositions or Test Papers Therefor; Processes of Preparing Such Compositions; Condition-Responsive Control in Microbiological or Enzymological Processes;
5	A61P	Specific Therapeutic Activity of Chemical Compounds or Medicinal Preparations;
6	A01K	Animal Husbandry; Care of Birds, Fishes, Insects; Fishing; Rearing or Breeding Animals, Not Otherwise Provided for; New Breeds of Animals;
7	C12P	Fermentation or Enzyme-Using Processes to Synthesise a Desired Chemical Compound or Composition or to Separate Optical Isomers from a Racemic Mixture;
8	G01N	Investigating or Analyzing Materials by Determining Their Chemical or Physical Properties;
9	C07H	Sugars; Derivatives Thereof; Nucleosides; Nucleotides; Nucleic Acids;
10	A01H	New Plants or Processes for Obtaining Them; Plant Reproduction by Tissue Culture Techniques.