doi:10.3772/j.issn.1006-6748.2021.02.013

Analysis on the development status of DNA storage based on papers and patents[①]

Zheng Yanning (郑彦宁), Yuan Fang[②], An Shudi
(Institute of Scientific and Technical Information of China, Beijing 100038, P. R. China)

Abstract

DNA storage as a disruptive technology is expected to solve the problem of massive data storage. Based on bibliometric analysis of DNA storage related papers and patents, this paper analyzes the development trend of DNA storage technology. The results show that DNA technology is still in the development stage, in which only a small number of researchers are involved. USA is the global leading country of DNA storage research. Both universities and companies in USA have played an important role in promoting DNA storage research. China is second only to USA in the number of DNA storage related papers or patents. However, in terms of patents layout, Chinese institutions don't have sufficient intention of opening up the global market in the application of DNA storage. Although there have been several breakthrough advances in DNA storage, there are still many challenges to be solved.

Key words: DNA storage, information, digital data, bibliometric analysis

0 Introduction

The digital revolution triggered a paradigm shift in how to generate and store information, resulting that digital data production has been growing exponentially. The total digital information today amounts to 3.52×10^{22} bits globally, and its consistent exponential rate of growth is expected to reach 3×10^{24} bits by 2040[1-2]. This trend will soon exceed the capacity of existing storage media, such as hard disk and flash memory. The current storage media have gradually exposed many problems. In recent years, deoxyribonucleic acid (DNA) as nature's oldest storage medium with the largest capacity has attracted widespread attention and has become an attractive potential data storage medium. In contrast to current storage media, longevity and stability are the significant advantages of DNA storage[3-4]. DNA molecules are very stable and easy to preserve without the harsh requirements of constant temperature and humidity. DNA molecules have super anti-electromagnetic interference ability. DNA data storage does not need to dissipate heat and it is not affected by physical factors such as vibrations, high temperatures, and magnetic fields. DNA molecules can endure over millennia under appropriate storage conditions[5-7]. DNA storage also has high volume and high density storage capacity.

DNA storage involves four major steps. First, pictures, videos, documents, and other information must be converted into a sequence of four bases in the DNA. Generally, the current storage media uses the stream of binary data to store information. In a similar way, DNA as an organic molecule consisting of four bases (A, C, G, T) is as a contiguous string of quaternary (base-4) numerals. Different encoding and compression algorithms are used in DNA storage. For binary data, the most intuitive conversion is representing 2 bits with one base (for example, A, C = 0; G, T = 1)[8-9]. However, the low information density prevented its use in later studies. Ternary transcoding methods are developed in DNA storage. A Huffman code translates binary to ternary digits, and a rotating encoding translates ternary digits to nucleotides[10]. With the continuous improvement of DNA storage encoding technology, many encoding methods such as exclusive-or (XOR) encoding have been developed[11-12]. Second, synthesize DNA (write) and store in vivo such as biological cells, bacteria or other living organisms[13-14], or store in vitro such as being frozen in solution or dried down for protection from the environment[15]. Ref. [16] devised a DNA-of-things (DoT) storage architecture to produce materials with immutable memory. The biggest breakthrough of this research

① Supported by the National Key Research and Development Program of China (No. 2019YFA0707203).
② To whom correspondence should be addressed. E-mail: yuanf@istic.ac.cn
Received on Feb. 3, 2021

is to confirm the theory that everything can realize DNA storage, and it is not restricted by any shape, which also makes the commercial value of DNA storage. Third, retrieve DNA data from the DNA pool in a process called random access. Random access within DNA data pools can be accomplished via polymerase chain reaction (PCR) based enrichment with primer pairs that map to specific data items generated during the encoding process. It is challenging for random access of DNA storage because of the lack of physical organization across data items in the same DNA pool. Magnetic bead and other methods can be supported in the random access processes of DNA storage[17-18]. Fourth, sequence the selected DNA (read) and decode DNA sequences back into the original digital data. The most common sequencing methods have low-throughput sequencing (for example, Sanger), high-throughput sequencing (for example, Illumina), nanopore sequencing for real-time data reading (for example, Oxford Nanopore Technologies), and so on[19].

1 Method and data collections

DNA storage has recently emerged as a promising approach for massive data storage because of its endurance, higher degree of compaction and unique biological properties. It is important for studying the development trend of DNA storage. At present, the study on development trend of DNA storage is mostly based on the method of literature review. From literature review, the knowledge of DNA storage history is more apparent, which is lacking of factual and objective data to support. DNA storage is an emerging technology, and its research enthusiasm is far less than the other emerging technologies, such as artificial intelligence or gene editing. Ref. [20] compared the differences between China and the United States in the field of DNA storage based on papers and patents. There are very little researches on DNA storage based on the bibliometric methods. Bibliometric analysis is a basic but an effective way to detect and examine the emergence of a new technology[21]. In this work, a bibliometric method based on papers and patents is used to analyse the development trend of DNA storage. Origin is used for drawing the statistical graphs. CiteSpace as a scientific visualization tool is used in the analysis of keywords. CiteSpace is used to deal with the data cleaning and the data preprocessing. Artificial analysis is used in keywords screening and organizations cleaning.

The papers data are retrieved from Web of Science Core Collection database. Search string is TS (Topic) = ((((store or storage) NEAR/1 (data or information or digital)) NEAR/5 DNA) or ('DNA-based storage') or ('DNA storage') or ((('nucleic acid $') or DNA) and (('information storage') or ('data storage'))) or ('DNA based memory') or ('DNA based memories') or ('message in DNA') or ('message into DNA') or ('messages in DNA') or ('messages into DNA') or ('DNA databases for the storage') or ('data storage' in 'nucleic acid $') or ('information based on DNA') or ('DNA memory') or ('signatures into genomic DNA') or ('digitized genome sequence information') or ('digital information in DNA') or ('Nucleic acid memory') or ('memory using the DNA')). Time span is set 'from 1900 to 2020'. 1,040 papers were obtained on January 7, 2021.

The patents data are retrieved from Innograpghy database. Search string is @(abstract, pclaims, title) (('DNA storage') or ('DNA based memory') or ('store data in DNA') or ('store information in DNA') or ('store digital information in DNA') or ('DNA data storage') or ('information storage with DNA') or ('data storage with DNA') or ('DNA-based storage') or ('data storage in nucleic acid') or ('DNA-based data storage') or ('information based on DNA') or ('digital information in nucleic acid') or ('digital information in DNA') or ('nucleic acid memory')) or @(abstract, title)(((('nucleic acid') or DNA) and (('information storage') or ('data storage'))). Filing time span is set 'from 1900-01-01 to 2020-12-31'. 569 patents were obtained on January 7, 2021.

Due to the lag of data statistics of Web of Science and Innography database, the data of 2020 may be incomplete and just for reference. As patents are affected by the examination system, there is generally a delay of 2 – 3 years from application to publication.

2 Results and analysis

2.1 The number of annual papers and patents distribution analysis

The number of annual papers and patents distribution (Fig. 1) shows a preliminary understanding of the development process of DNA storage. Before 1990, the study of DNA storage appeared to be piecemeal. There was a small peak in the study of DNA storage between 2001 and 2007, of which 54 patents were filed in 2005. After 2015, the number of papers and patents was increasing year by year. The number of annual papers peaked at 119 in 2019, and the number of annual patents peaked at 87 in 2018. The basic concept of using DNA for data storage can be dated back to the mid-1960s, when Refs[22-23] discussed the idea of genetic memory. Since then, many researchers have worked on developing this initial idea towards a viable product. Ref. [24] demonstrated microvenus icon coding in synthetic DNA in 1988 and an ancient Germanic rune was incorporated into live Escherichia coli cells in 1996. In the early 2000s, a simple way was proposed to use codon triplets for encoding alphabets, suggesting great potential for DNA as a storage medium[25]. A major breakthrough occurred in the early 2010s, which

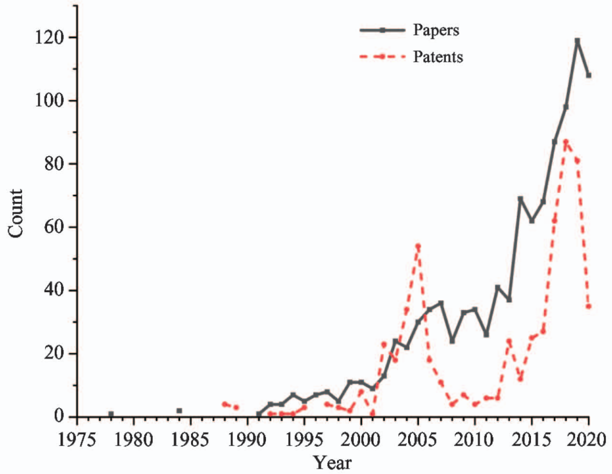

Fig. 1 Number of annual papers and patents distribution

is the feature key work on the modern reincarnation and demonstration of DNA storage[8,10]. Recently, the amounts of DNA storage research are gradually increasing with increasingly complex methods, a holistic, novel solution for data handling completely run on DNA could become a reality.

2.2 The subject category distribution analysis

Table 1 and Fig. 2 show the subject category distribution of DNA storage. The subject categories of DNA storage related papers are classified according to the function of Web of Science categories. Table 1 shows the top 15 Web of Science categories of DNA storage related papers. The top 5 categories of DNA storage related papers are engineering electrical electronic, chemistry multidisciplinary, computer science theory methods, multidisciplinary sciences and biochemistry molecular biology. Each of the top 5 categories has more than 100 related papers. Nano science, materials science and physics are also shown in the lower rank categories. The Cooperative Patent Classification (CPC) is a patent classification system, which has been jointly developed by the European Patent Office (EPO) and the United States Patent and Trademark Office (USPTO).

Table 1 Web of Science categories of DNA storage related papers

Rank	Categories	Count	Proportion/%
1	Engineering electrical electronic	131	12.596
2	Chemistry multidisciplinary	115	11.058
3	Computer science theory methods	107	10.288
4	Multidisciplinary sciences	106	10.192
5	Biochemistry molecular biology	101	9.712
6	Computer science information systems	94	9.038
7	Biotechnology applied microbiology	80	7.692
8	Biochemical research methods	75	7.212
9	Nanoscience nanotechnology	71	6.827
10	Materials science multidisciplinary	71	6.827
11	Mathematical computational biology	61	5.865
12	Genetics heredity	60	5.769
13	Physics applied	55	5.288
14	Chemistry physical	48	4.615
15	Computer science artificial intelligence	46	4.423

CPC is substantially based on the previous European classification system (ECLA), which itself is a more specific and detailed version of the International Patent Classification (IPC) system. CPC class of DNA storage related patents is classified according to CPC class function of Innography. There are 6 CPC main groups of DNA storage related patents in Fig. 2, C (chemistry metallurgy), G (physics), B (performing operations transporting), A (human necessities), H (electricity) and F (mechanical engineering lighting heating weapons blasting). The top 5 secondary classifications are C12 (biochemistry beer spirits wine vinegar microbiology enzymology mutation or genetic engineering), G06 (computing calculating counting), B01 (separating mixing), B82 (nanotechnology) and G16 (B0/00: …). Based on the results of Table 1 and Fig. 2, DNA storage technology covers a variety of subject category, such as chemistry, biology, physics, computer science, nano science and materials science, which is a multidisciplinary technology.

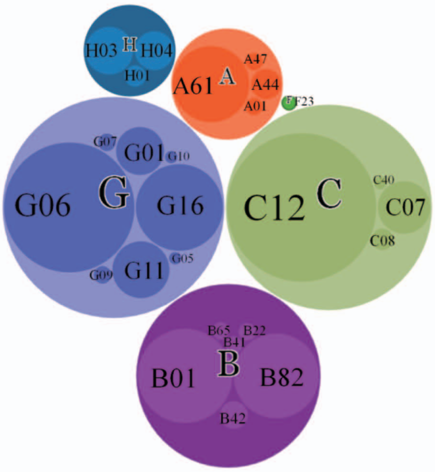

Fig. 2 CPC class of DNA storage related patents

2.3 The country distribution analysis

In the field of DNA storage, a total of 69 countries or regions have published relevant papers and the top 15 countries or regions of DNA storage related papers are shown in Fig. 3. The number of DNA storage related papers in USA is 438, accounting for 42.115% of the world's total. USA is the main country for DNA storage related papers published and China is close behind. The number of DNA storage related papers in China is 131, accounting for 12.596% of the world's total. The number of DNA storage related papers of these two countries accounts for more than half of the world's total. England, Germany and France rank third to fifth respectively and each country accounts for more than 5% of the word's total. The number of DNA storage related papers in other countries is less than 50, accounting for less than 5% of the word's total.

In the field of DNA storage, a total of 19 countries' inventors filed relevant patents and the source jurisdictions are a total of 57 countries or organizations. The top 9 locations of DNA storage related patents' inventor and the top 10 source jurisdictions of these patents are shown in Fig. 4. The number of source jurisdictions counts expand due to EPO documents included in each active designated state. Similar to the result of Fig. 3, USA is the main country for DNA storage related patents filed and China is also close behind. The number of DNA storage related patents in the location of USA is 266 and in the source jurisdiction of USA is 139. In these 266 patents, there are 104 patents, whose source jurisdiction is USA. DNA storage related patents' inventors in USA are inclined to develop a global patent layout. The number of DNA storage related patents in the location of China is 105 and in the source jurisdiction of China is 116. In these 105 patents, there are 90 patents, whose source jurisdiction is China. Patents inventors will apply for patents in different target market countries to achieve the purpose of occupying the international market. The source jurisdictions are in more countries, the greater the technical value or market value of the patents are. By comparison, USA inventors pay more attention to the application market of DNA storage.

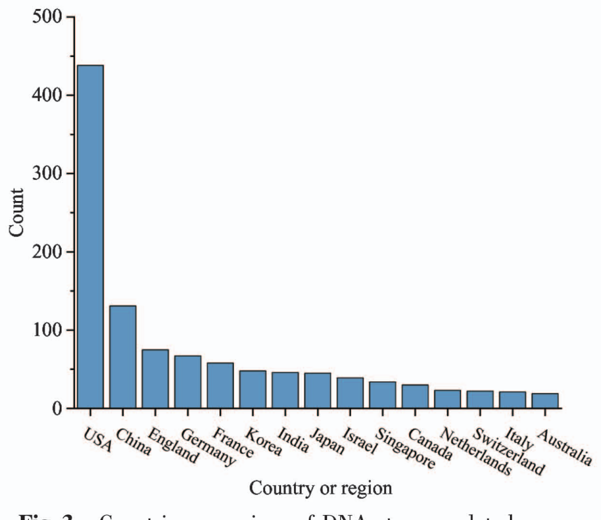

Fig. 3 Countries or regions of DNA storage related papers in the world

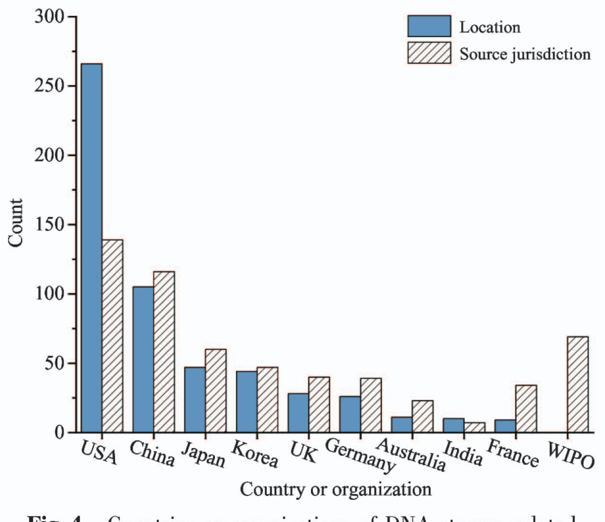

Fig. 4 Countries or organizations of DNA storage related patents in the world

2.4 The institutions distribution analysis

In the field of DNA storage, there are a total of 163 institutions filed relevant patents and a total of 1205 institutions published relevant papers. The top 15 institutions of DNA storage related papers and patents are shown in Table 2. In the top 15 institutions of DNA storage related papers, Microsoft Corporation is the only one company. Universities are the major participants in the publication of DNA storage related papers. University of California System, French National Centre for Scientific Research, and University of Illinois System ranked first to third respectively in the key institutions of papers. In the top 15 institutions of DNA storage related patents, there are 10 companies, which indicates that companies are the main participants in the applications of DNA storage related patents. Twist Bioscience Corporation is a rapidly growing synthetic biology company that has developed a disruptive DNA synthesis platform to industrialize the engineering of biology. Catalog Technologies is building the world's first DNA-based platform for massive digital data storage and computation. Roswell Biotechnologies is a molecular electronics company transforming DNA sequencing with their latest platform, designs to deliver the disruptive performance required for future biomedical and industrial applications. In the field of DNA storage, its participating companies mainly include DNA synthesis companies, DNA sequencing companies, and computer software development companies and so on.

Table 2 Key institutions in the field of DNA storage

Rank	Institutions of papers	Count	Location	Institutions of patents	Count	Location
1	University of California System	51	USA	Twist Bioscience Corporation	31	USA
2	The French National Centre for Scientific Research	35	France	European Molecular Biology Laboratory	31	Europe
3	University of Illinois System	31	USA	Catalog Technologies	27	USA
4	Nanyang Technological University	21	Singapore	Roswell Biotechnologies	17	USA
5	Chinese Academy of Sciences	20	China	Harvard University	15	USA
6	Stanford University	20	USA	Konica Minolta	12	Japan
7	University of Washington	20	USA	Technicolor	10	France
8	Harvard University	19	USA	Chinese Academy of Sciences	9	China
9	University System of Georgia	18	USA	Unilever	9	England
10	Microsoft Corporation	17	USA	Microsoft Corporation	9	USA
11	Israel Institute of Technology	17	Israel	Dodo Omnidata	9	USA
12	University of North Carolina	15	USA	Massachusetts Institute of Technology	9	USA
13	Swiss Federal Institute of Technology Zurich	14	Switzerland	Molecular Assemblies	8	USA
14	National University of Singapore	14	Singapore	Tianjin University	8	China
15	United States Department of Energy	14	USA	Illumina	8	USA

2.4.1 The time-zone of keywords analysis

By plotting the time zone of keywords, the dynamic development process of the gene editing research can be obtained from Fig. 5, which describes in Citespace with the paper data. DNA storage research originated in the early 1990s. In the 1990s, DNA storage research was scarce, which keywords are 'databese', 'sequence', 'gene' and so on. From 2000 to 2011, more and more researches began in the field of DNA storage. In this stage, researchers began to try various methods to store information in DNA, which keywords are 'computation', 'storage', 'nanostructure', 'algorithm' and so on. However, these early attempts only stored less than tens of Bytes with little scalability for practical usages. It was not until the early 2010s that the groundbreaking works of Refs [8,10] led to the return of DNA storage to mainstream interest. After 2011, complex compilation methods are increasing to solve major challenges in DNA storage research and applications, for example, the 'random access'. The first robust system based on error correcting codes using inner codes and outer codes for DNA data storage was developed in 2015[26], and it demonstrates silica encapsulation for greater durability. Therefore, 'synthetic biology', 'code' and 'robust' as the keywords emerged in this stage.

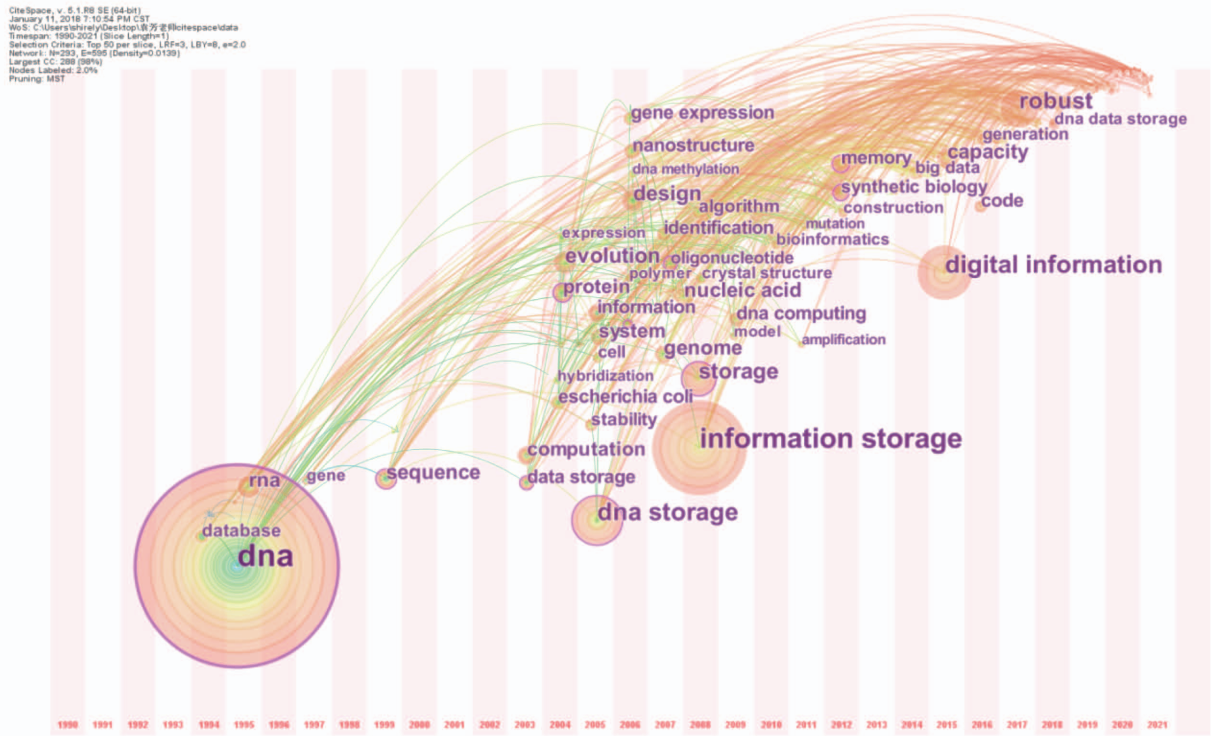

Fig. 5 The time-zone of keywords in the field of DNA storage

2.4.2 Keyword clustering analysis

Through keyword clustering network analysis of the DNA storage related papers, 11 keyword clusters shown in Fig. 6 are calculated using log-likelihood ratio (LLR) cluster labeling extraction algorithms by CiteSpace: 'big data', 'DNA data', 'storage', 'mechanism', 'repetitive element', 'error correction codes', 'DNA nanotechnology', 'systems biology', 'polymer', 'electrical transport', 'alphabet' and 'telomere'. The value of silhouette for each cluster is greater than 0.78, indicating robust and meaningful results. The lowest silhouette value of 0.789 is the Cluster #4 (error correction codes) with 29 members and the other silhouette values are higher than 0.800. DNA synthesis and sequencing are error prone. Error correction codes boil down to adding redundant information, which increases the probability that the original information can be retrieved even in the presence of errors or missing data[19]. Multiple research projects have developed coding for DNA data storage to cope with read and write errors and with degradation of DNA over time[11,17,26-28]. The largest cluster is Cluster #0 (big data) which has 51 members and a silhouette score of 0.809. DNA storage offers density of up to 10^{19} bits/cm^3, several orders of magnitude denser than the densest media available today. For example, the DNA storage device in our palms size can accommodate the information of the entire world[1]. It is great potential for DNA storage in solving massive data storage in the future. Cluster #7 (polymer) relates to the PCR process, which involves in the replication and amplification of DNA. Cluster #9 (alphabet) relates to the DNA bases and genetic alphabet.

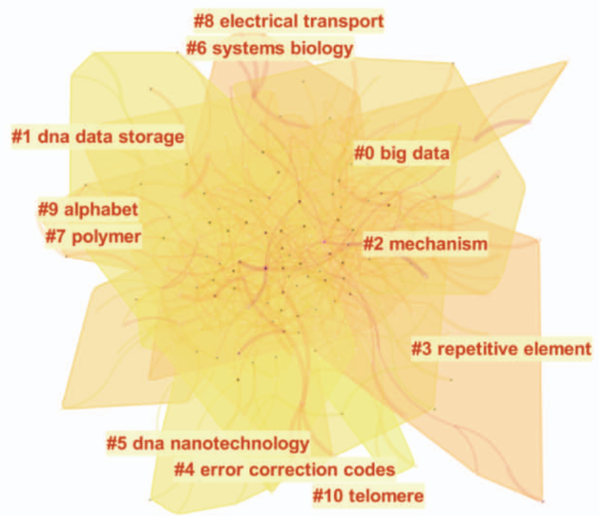

Fig. 6 Keyword clustering network analysis for DNA storage

Fig. 7 is the keyword clustering of DNA storage related patents, which is drawn by the PatentScape function of Innography, 20 hot topics are shown in Fig. 7. These 20 topics are related to base, sequence, and storage. These are the most closely related topics in DNA storage technology. 'Data storage', 'information storage' and 'DNA storage' ranked third to fifth, with 209 patents, 113 patents and 102 patents respectively. Compared to the result of the keyword clustering of DNA storage related papers, these topics in Fig. 7 are more biased towards practical applications and lack basic research topics, such as 'mechanism', 'codes', 'model', 'design' and so on.

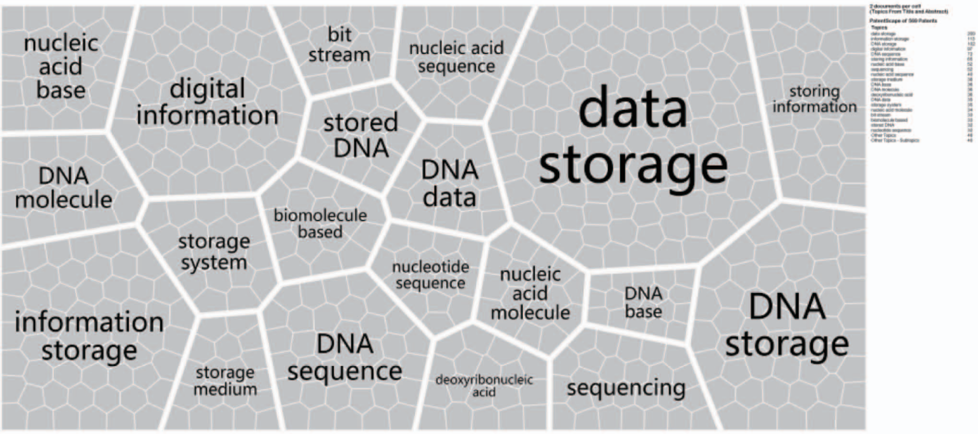

Fig. 7 Details of cluster DNA storage related patents

3 Conclusions

DNA storage is an emerging technology with considerable potential to be the next generation storage medium of choice. Compared with current storage media, DNA storage has the advantages of super storage capacity, stability and safety. In the future, information that is not commonly used but needs to be preserved, such as government documents, historical archives, etc., are especially suitable for DNA storage. USA is the global leading DNA storage research country. USA attaches great importance to DNA storage research and has formulated many related research projects since 2017. With the strong support of the government, both universities and companies in USA have played an important role in promoting DNA storage research. In particular, the international giant Microsoft Corporation has made many contributions to the industrialization of DNA storage. In 2019, Microsoft Corporation developed an automated end-to-end DNA data storage device to explore the challenges of automation within the constraints of this unique application[29]. Microsoft Corporation realized the world's first fully automatic DNA data writing and reading device, which is a key step for DNA storage technology from the laboratory to commercial applications. China is the second only to USA in the number of DNA storage related papers or patents. However, in terms of patents layout, Chinese institutions don't have sufficient intention of opening up the global market in the application of DNA storage. At present, DNA technology is still in the developing stage. Although there have been several breakthrough advances in DNA storage, there are still many challenges to be solved. For example, cost, read speed, random access and automation have become key issues limiting the industrialization of DNA storage. In view of the current situation that China needs to strengthen DNA storage research, and this work recommends that China should make a strategic plan for DNA storage technology. First, financial support and relevant supporting policies of development plans for DNA storage should be strengthened. Second, it is very important to pay attention to the application of DNA storage technology in national security. Military-civilian integration project research of DNA storage should be strengthened to support the potential applications of DNA storage in national security and defense. Third, DNA storage technology integrates multiple disciplines, and it requires the cross cooperation of multidisciplinary researchers. Therefore, universities and other research institutions should be encouraged to build an interdisciplinary research team for biology and computer information science to promote the exchange of multidisciplinary researchers and strengthen interdisciplinary basic research.

Reference

[1] Panda D, Molla K A, Baig M J, et al. DNA as a digital information storage device: hope or hype? [J]. 3 *Biotech*, 2018, 8(5): 1-9

[2] Zhirnov V, Zadegan R M, Sandhu G S, et al. Nucleic acid memory[J]. *Nature Materials*, 2016, 15(4): 366-370

[3] Dietrich A, Been W. Memory and DNA[J]. *Journal of Theoretical Biology*, 2001, 208: 145-149

[4] Garzon M H, Neel A, Chen H. Efficiency and reliability

of DNA based memories[J]. *GECCO*, 2003, 2723: 379-389

[5] Allentoft M E, Collins M, Harker D, et al. The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils[J]. *Proceedings of the Royal Society B: Biological Sciences*, 2012, 279(1748): 4724-4733

[6] Branzei D, Foiani M. Regulation of DNA repair throughout the cell cycle[J]. *Nature Reviews Molecular Cell Biology*, 2008, 9(4): 297-308

[7] Stanley P M, Vickers A M, Strittmatter L M, et al. Decoding DNA data storage for investment[J]. *Biotechnology Advances*, 2020, 45: 107639

[8] Church G M, Gao Y, Kosuri S. Next-generation digital information storage in DNA[J]. *Science*, 2012, 337(6102): 1628

[9] Reed I S, Solomon G. Polynomial codes over certain finite fields[J]. *Journal of the Society for Industrial and Applied Mathematics*, 1960, 8(2): 300-304

[10] Goldman N, Bertone P, Chen S, et al. Towards practical, high-capacity, low-maintenance information storage in synthesized DNA[J]. *Nature*, 2013, 494(7435): 77-80

[11] Organick L, Ang S D, Chen Y J, et al. Random access in largescale DNA data storage[J]. *Nature Biotechnology*, 2018, 36(3): 242-248

[12] Shendure J, Balasubramanian S, Church G M, et al. DNA sequencing at 40: past, present and future[J]. *Nature*, 2017, 550(7676): 345-353

[13] Lee H, Popodi E, Tang H X, et al. Rate and molecular spectrum of spontaneous mutations in the bacterium Escherichia coli as determined by whole-genome sequencing[J]. *Proceedings of the National Academy of Sciences*, 2012, 109(41): E2774-E2783

[14] Shipman S L, Nivala J, Macklis J D, et al. CRISPR-Cas encoding of a digital movie into the genomes of a population of living bacteria[J]. *Nature*, 2017, 547(7663): 345-349

[15] Organick L, et al. Random access in large-scale DNA data storage[J]. *Nature Biotechnology*,2018, 36(3): 242-248

[16] Koch J, Gantenbein S, Masania K, et al. A DNA-of-things storage architecture to create materials with embedded memory[J]. *Nature Biotechnology*, 2020, 38(1): 39-43

[17] Yazdi S M H T, Yuan Y, Ma J, et al. A rewritable, random-access DNA based storage system[J]. *Scientific Reports*, 2015, 5: 14138

[18] Organick L, Ang S D, Chen Y J, et al. Random access in large-scale DNA data storage[J]. *Nature Biotechnology*, 2018, 36(3): 242-248

[19] Ceze L, Nivala J, Strauss K. Molecular digital data storage using DNA[J]. *Nature Reviews Genetics*. 2019, 20(9): 456-466

[20] Yuan F, Zheng Y N, Zheng J, et al. A comparative study between China and the United States in the field of DNA storage technology[J]. *Global Science, Technology and Economy Outlook*, 2019, 34(4): 71-76 (In Chinese)

[21] Braun T, Schubert A. A quantitative view on the coming of age of Interdisciplinarity in the sciences, 1980-1999[J]. *Scientometrics*, 2003, 58(1): 183-189

[22] Wiener N. Interview: machines smarter than men? [J]. *U. S. News and World Report*, 1964, 56: 84-86

[23] Neiman M S. On the molecular memory systems and the directed mutations[J]. *Radiotekhnika*, 1965, 6: 1-8

[24] Davis J. Microvenus[J]. *Art Journal*, 1996, 55:70-74

[25] Bancroft C, Bowler T, Bloom B, et al. Long-term storage of information in DNA[J]. *Science*, 2001, 293(5536): 1763-1765

[26] Grass R N, Heckel R, Puddu M, et al. Robust chemical preservation of digital information on DNA in silica with error-correcting codes[J]. *Angewandte Chemie International Edition*, 2015, 54(8): 2552-2555

[27] Heckel R, Shomorony I, Ramchandran K, et al. Fundamental limits of DNA storage systems[C]//IEEE International Symposium on Information Theory, Aachen, Germany, 2017: 3130-3134

[28] Erlich Y, Zielinski D. DNA Fountain enables a robust and efficient storage architecture[J]. *Science*, 2017, 355(6328): 950-954

[29] Takahashi C N, Nguyen B H, Strauss K, et al. Demonstration of end-to-end automation of DNA data storage[J]. *Scientific Reports*, 2019, 9(1): 4998

Zheng Yanning, born in 1965. He is a professor of information science and a postdoctoral supervisor. His research interests include theory and methods of information science.