DNA DATA STORAGE RESEARCH

DNA data storage is an emerging field that leverages the unique properties of DNA molecules to address the growing demand for data storage, offering high density, durability, and energy efficiency.

Overview of DNA Data Storage

DNA data storage involves encoding digital information into the sequences of DNA molecules. This method is being explored as a solution to the limitations of traditional storage media, which are struggling to keep pace with the exponential growth of data. For instance, DNA can theoretically store up to **215 petabytes** (215 million gigabytes) of data in just one gram, making it an incredibly compact storage medium.

Current Research and Developments

1. **Collaborative Efforts**: Microsoft and the University of Washington are at the forefront of DNA data storage research, focusing on developing methods to synthesize and manipulate DNA for archival storage. Their work aims to create a high-density, durable, and easy-to-manipulate storage medium.
2. **Technical Challenges**: Despite its potential, DNA data storage faces several challenges, including the high costs of DNA synthesis and sequencing, which were estimated at about **$3,500 per gigabyte** in 2020. Additionally, the speed of writing and reading data in DNA is significantly slower than traditional electronic storage.
3. **Advancements in Encoding**: Researchers have successfully encoded substantial amounts of data into DNA, including the entire **16 GB text of the English Wikipedia**. Innovations in DNA synthesis technology are improving the efficiency and speed of data encoding and retrieval.
4. **Future Prospects**: The demand for data storage is projected to reach **175 zettabytes** by 2025, highlighting the urgent need for new storage solutions. DNA storage is seen as a viable option due to its

longevity (with a half-life of over **500 years**) and stability under proper conditions.

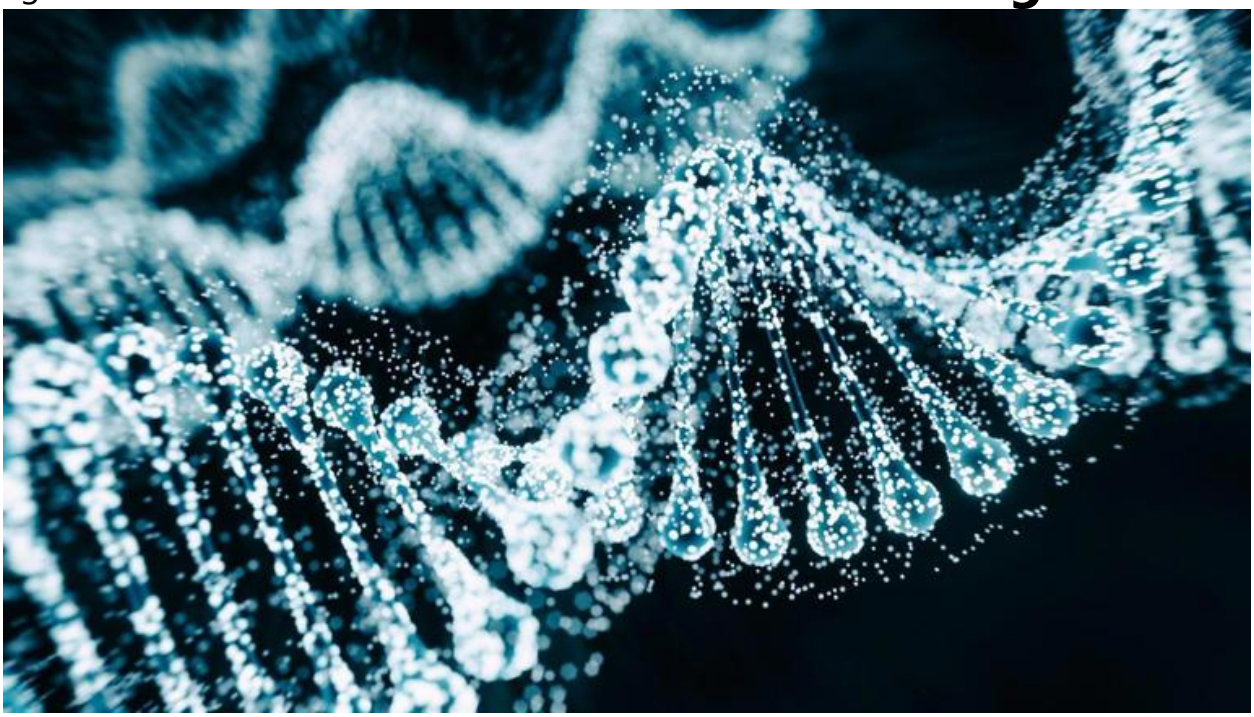
Conclusion

As the field of DNA data storage continues to evolve, it holds promise for addressing the challenges posed by the ever-increasing volume of digital data. Ongoing research aims to overcome current limitations, making DNA a practical and sustainable solution for future data storage needs. The integration of biotechnology with data storage technology could revolutionize how we think about and manage information in the digital age.

DNA able to store 1TB of data in a drop

DNA has been recognized not only as the blueprint of life but also as a potential medium for digital data storage. With recent breakthroughs, scientists have demonstrated that DNA can store up to 1 terabyte (TB) of data in just a single droplet. The intersection of biology and data science presents a revolutionary approach to data storage, poised to address the growing needs of the digital age.

The Science Behind DNA Data Storage



Structure and Encoding

The double-helix structure of DNA, composed of sequences of nucleotides—adenine (A), thymine (T), cytosine (C), and guanine (G)—offers a unique way to encode digital information. Each nucleotide can represent binary data, allowing DNA to serve as a high-density storage medium. By assigning binary values to these nucleotides, scientists can translate the endless stream of zeros and ones into sequences of A, T, C, and G, effectively turning DNA into a biological hard drive.

Encoding Process

The process of encoding involves converting binary data into DNA sequences through algorithms that ensure efficient mapping. DNA synthesis technology then creates these sequences, constructing physical strands of DNA that embody the digital data. This synthesis is followed by sequencing, a step that enables the reading of encoded data. According to a study by MIT, advances in sequencing technologies play a pivotal role in accessing stored information, further bridging the gap between biological systems and digital data.

Error Correction

Ensuring data integrity in DNA storage involves sophisticated error correction algorithms. These algorithms detect and correct errors that may occur during synthesis or sequencing. Redundancy is also built into the system, where multiple copies of data sequences are stored. This redundancy, combined with error correction techniques, provides a robust framework for maintaining the accuracy and reliability of stored data, as highlighted in a research article.

Advantages Over Traditional Storage Media

Density and Capacity

Traditional storage devices such as hard drives and SSDs are rapidly reaching their physical limitations. In contrast, DNA offers an unprecedented data density, capable of storing vast amounts of information in a minuscule space. A single gram of DNA can theoretically hold up to 215 petabytes of data, a significant leap from current storage technologies. This incredible density is a game-changer for industries dealing with massive data sets, opening up new possibilities for data management and storage solutions.

Durability and Longevity

DNA's stability is another compelling advantage. Unlike electronic storage media, which can degrade over time and are susceptible to environmental damage, DNA is incredibly durable. Fossilized DNA found to be thousands of years old retains its structure, demonstrating its resilience. This durability means that DNA-stored data could potentially last for centuries, providing a reliable solution for long-term data archiving.

Energy Efficiency

The energy efficiency of DNA data storage is particularly appealing in an era where environmental sustainability is paramount. Traditional data centers consume vast amounts of energy to maintain electronic storage systems. DNA, however, requires minimal energy for data maintenance, reducing the carbon footprint associated with data storage. This energy efficiency aligns with global efforts to create more sustainable technological infrastructures.

Challenges and Limitations

Cost and Scalability

While DNA data storage offers numerous advantages, the technology is not without its challenges. The cost of DNA synthesis and sequencing remains high, posing a significant barrier to widespread adoption. Scaling the technology to make it cost-effective for everyday use is a major hurdle that researchers and companies must overcome. The current research focuses on reducing these costs

through technological advancements and increased efficiency in synthesis processes.

Read/Write Speed

Another limitation is the read and write speeds of DNA data storage systems. The process of encoding and retrieving data from DNA is time-intensive, far slower than electronic systems. Improving these speeds is crucial for the practical application of DNA storage, requiring innovations in both synthesis and sequencing technologies. As noted in a recent study, developing faster methods for data retrieval is essential for DNA storage to compete with traditional systems.

Technical Complexities

The technical complexities involved in DNA data storage are considerable. Precise handling of DNA molecules is essential to prevent errors in encoding and decoding processes. This precision demands sophisticated equipment and expertise, adding layers of complexity to the implementation of DNA storage systems. Overcoming these challenges is vital for the technology to reach its full potential, making it accessible and practical for a wider range of applications.

Real-World Applications and Future Prospects

Image Credit: United States Mission Geneva – CC BY 2.0/Wiki Commons© Image Credit: United States Mission Geneva - CC BY 2.0/Wiki Commons

Data Archiving

One of the most promising applications of DNA data storage is in long-term data archiving. Enterprises, research institutions, and governmental bodies can benefit from the technology's durability and density for storing critical information. DNA's ability to preserve data without degradation makes it an ideal solution for archiving historical records, scientific data, and other valuable information.

Biocompatible Devices

DNA data storage also holds potential for the development of biocompatible devices. These devices could integrate biological systems with digital technology, opening up new avenues for biomedical applications. For instance, implantable

devices that store patient data in DNA form could revolutionize healthcare by providing continuous monitoring and personalized treatment options.

Future Innovations

The future of DNA data storage is ripe with possibilities. Ongoing research aims to address current limitations, such as cost and speed, making the technology more accessible and practical. Potential breakthroughs could include the development of automated synthesis and sequencing systems, further bridging the gap between biological and digital domains. The future may also see the integration of DNA storage with emerging technologies like quantum computing, enhancing its capabilities and applications.

Ethical and Environmental Considerations

Ethical Implications

As with any emerging technology, DNA data storage raises ethical questions. The manipulation of DNA for non-biological purposes prompts concerns about the potential misuse of the technology and its impact on data privacy. These ethical implications require careful consideration and regulation to ensure that DNA storage is used responsibly and ethically.

Environmental Impact

From an environmental perspective, DNA data storage offers significant benefits. By reducing the reliance on electronic storage systems, DNA storage can decrease electronic waste and the carbon footprint of data centers. This aligns with efforts to create more sustainable technological solutions, underscoring the potential of DNA storage to contribute positively to environmental conservation.

DNA could revolutionize how we store our data

The Age of AI will rely on massive volumes of data that can be easily stored and retrieved—and bioscience may have an ingenious solution. DNA (deoxyribonucleic acid). August 21, 2025

Shakespeare's entire catalog of sonnets and eight of his tragedies, all of Wikipedia's English-language pages, and one of the first movies ever made: scientists have been able to fit the contents of all these works in a space smaller than a tiny test tube. They didn't somehow miniaturize them, though. Instead, they used DNA—the building block of all life—to encode the information in these creative works and store it at a microscopic scale.
As humans adopt advanced tools like artificial intelligence, tomorrow's currency will be data. Already, tech giants like Microsoft are raising billions of dollars to construct data centers for AI. And there's a very real "Storage Wars" scramble underway right now to figure out how to preserve and safeguard exponentially increasing amounts of data. Football field-size, gigawatt energy-sucking data centers are one option. Or DNA storage could be an energy-efficient, compact solution.

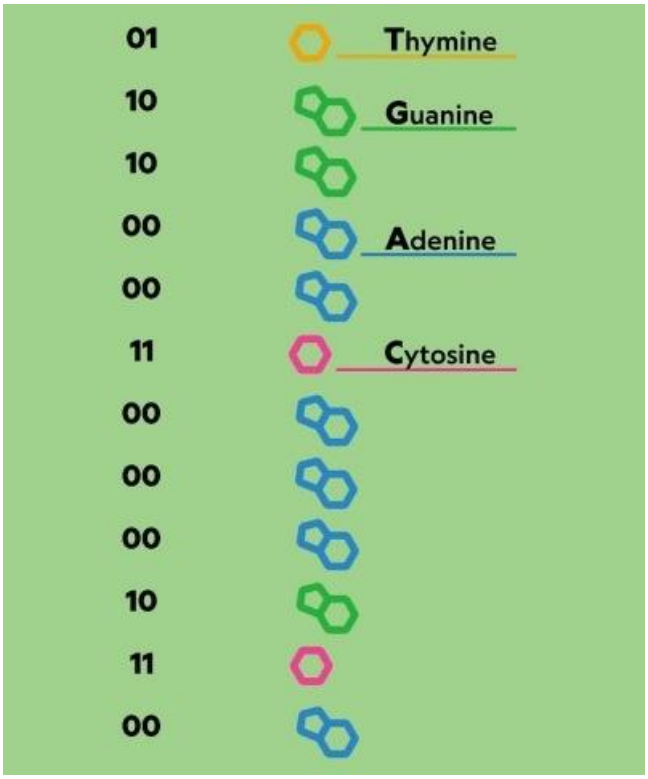
Step 1: Computer storage

We typically think of DNA as a blueprint or instruction booklet—its sequences of As, Ts, Cs, and Gs tell molecular machines how to build the fabric of our very beings. DNA storage flips this paradigm on its head. Computer data make up the inputs, and DNA is the end product. A handful of start-ups are working to perfect the conversion of binary computer code into physical DNA strands, and in doing so, take a shot at disrupting the multibillion-dollar storage industry. Here's how they plan to move the industry away from microfilm, microfiche, disks, and servers.

Traditional data storage relies on constant migration to prevent old data from degrading or the technology it's stored in from becoming obsolete. Varun Mehta,

CEO of Atlas Data Storage, compares long-term data storage to painting the Golden Gate bridge—by the time you've gone from one end to the other, the first end is rusting and you have to start all over again. "The same thing happens with long-term data storage," he says. "You're always moving from your old tape to your new tape." He predicts that "people who want to get off that treadmill will be the first to move to DNA."

Step 2: Encoding digital data in DNA



STEP 2: TO store digital information on DNA, algorithms convert digital codes into combinations of the four letters,

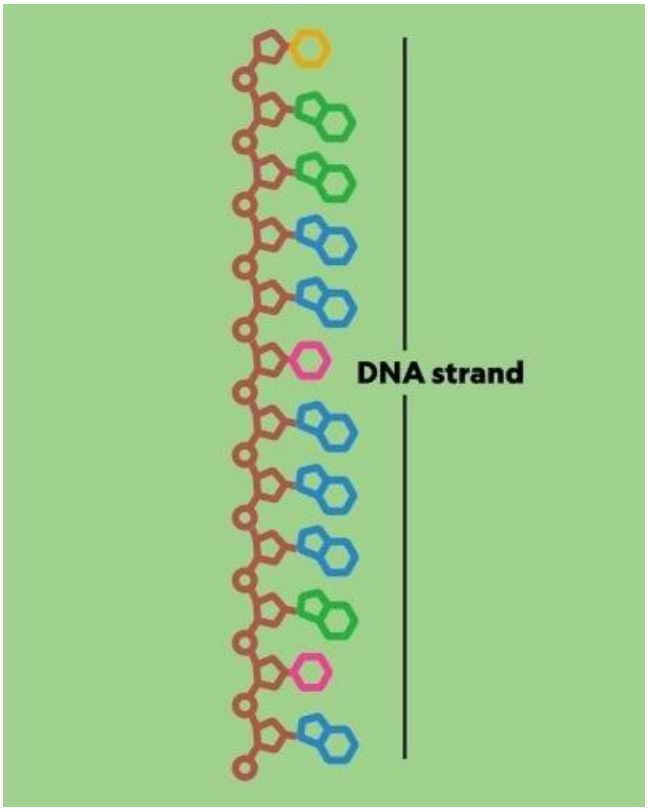
In practice, DNA storage involves several steps: deciding on a code, making the DNA using a process called synthesis, and storing the resulting DNA strands. DNA storage methods also include ways to categorize the stored strands and convert nucleotide sequences back into information that may be compatible with computers or accessible in some other way. Though industry members formed the DNA Data Storage Alliance in 2020 in part to set standards, companies in the DNA storage space still approach each of these steps in slightly different ways.

First, to store information as DNA, scientists have to determine how the data will be translated. DNA is a base 4 system; in contrast, computers store and process information in binary. Instead of assigning a “1” or a “0” to each DNA nucleotide—an A, C, T, or G—you could instead assign a particular combination of two digits to each base—so an A might stand in for “00,” C “01,” T “10,” and G “11.” Theoretically, this means every DNA nucleotide can encode up to 2 unique bits. In practice, the system isn’t as efficient as that (there are certain combinations of DNA nucleotides that are less stable or otherwise undesirable, and different chemistry protocols exist for turning bits into DNA bases).

Catalog, one DNA storage company, announced in 2022 that it had encoded eight of William Shakespeare’s tragedies into a single test tube. To do this, scientists had to translate about 207,000 words into strings of nucleotides using a class of enzymes called recombinases. They claimed their DNA-building machine, Shannon, encoded the plays into millions of nucleotides in a matter of minutes.

“To each of those words, you associate a random bit vector. A bit vector is just a sequences of ones and zeroes of a fixed length,” explains Catalog’s head of DNA Computing, Swapnil Bhatia, in a video for the company. The word “rose” might have a random bit vector stretching 1,000 numbers long, and different companies will have different ciphers for translating words into 1s, 0s, and nucleotides.

Step 3: Synthesis



STEP 3 These letters are one part of the nucleotides, the building blocks of DNA.

DNA synthesis—the step of actually creating custom DNA strands—is another place where companies diverge in their methods. Catalog uses the principles of inkjet printing to exude tiny droplets containing premade DNA fragments. In each droplet, hundreds of thousands of chemical reactions take place per second to elongate the DNA strands. Atlas Data Storage, meanwhile, relies on semiconductor chips and silicon wafers as the environment for assembling strands of synthetic DNA.

“Once those strands are assembled, we harvest them from our chip,” Mehta says. “These DNA strands really are like corn stalks growing in a field on this chip and once they've gotten to the height that we want—to the number of bases—then we harvest them.”

Step 4: Storing the DNA

STEP 4 The strands are archived in material like sealed vials or other materials.

Storing and preserving these synthetic strands presents another set of hurdles. Catalog and Atlas store DNA samples inside metal capsules, where the strands are not exposed to the elements and degraded. To convert DNA back into bit form, one can sequence it—using the same technology that powers genetic testing like 23andMe. This method can't be done indefinitely; eventually, the sample will need to be copied over again to restore it. To create longer-lasting, accessible storage, some groups are working on fluorescent tags. Shining a light on the samples can tell researchers information about a given sample at a glance, the same way metadata can help us organize computer files without having to open them.

If companies are able to surmount these challenges, a DNA storage system would take up a fraction of the space of traditional storage methods.

"The theoretical limit is astounding," Mehta says. "You could fill 50 petabytes worth of data in in a Tylenol-sized capsule"—or roughly 50,000 times as much data as an iPhone can store.

Step 5: Retrieving the data

STEP 5 When it's time to retrieve digital data from DNA, an algorithm converts the archived information back to...Read More

Storing information in such a small physical package raises philosophical questions about the purpose of storage. Could a storage device itself serve a purpose? Scientists have theorized and created proofs-of-concept of fabrics and everyday items like glasses that contain DNA-stored information. The company Catalog has a branch dedicated to "DNA computing" to search and analyze synthetic DNA without first converting the information encoded in it back into bits. There could be some advantages to working with data in DNA form—rather than moving from one end to another, like a computer processor does, working with the data can occur in many places at once in parallel.

DNA's status as the basic building block of life may someday make it one of our most durable technologies, Mehta says, because it means it isn't going anywhere.

"One thousand years from now, there probably will not be any DVD players. In fact, it's hard to find a VHS tape player anymore. But that's never going to happen with DNA, because we need it for our own health," he says. "We'll always have that technology available."

Twist Bioscience: A 'Hold' In DNA Data Storage Twist Bioscience Corporation (TWST) StockTWST

Myriam Alvarez

Summary

- Twist Bioscience specializes in synthetic biology and DNA applications, focusing on synthetic DNA, DNA products, and Next-Generation Sequencing tools.
- The company's revenue-generating segments, particularly DNA synthesis, show high growth and cost reduction potential.
- Financial discipline and strategic expansion show a path toward profitability, with a solid cash runway supporting near-term operations.
- Despite its promising business outlook, TWST's stock price is already inflated, leading to a "hold" rating, but I consider it a great buy on dips.

janiecbros

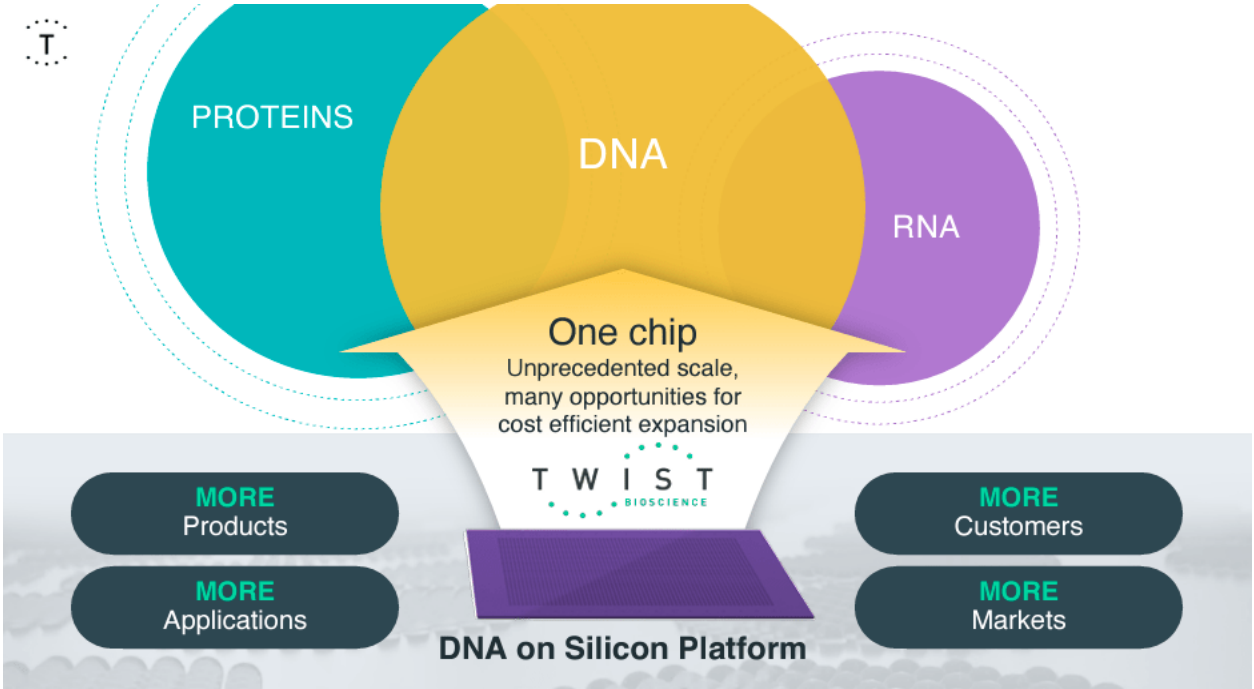
Twist Bioscience Corporation (NASDAQ:TWST) is a promising bet on DNA applications with synthetic DNA, DNA products, and Next-Generation Sequencing [NGS] tools for researchers in several industries. TWST's IP has applications in health care, materials

and chemicals, agriculture, and academic research. The company's Express Genes service effectively expanded TWST's customer base, and as a result, its Q1 2024 report came in above management's guidance. Yet, despite all these positives, my valuation analysis suggests TWST's stock price already reflects this potential, trading at a relatively inflated valuation. This tempers my optimism on the stock itself. Hence, I rate it a "hold," but I consider TWST a great buy on dip for opportunistic investors.

A Bet on DNA Applications: Business Overview

Twist Bioscience specializes in synthetic biology, delivering artificial DNA, DNA products, and NGS systems and tools. The company was founded in 2013 and is based in South San Francisco, California. TWST's platform automates and industrializes synthetic biology with a new method to produce artificial DNA by writing its information on a silicon chip, reducing production times and costs. You can think of TWST as a company that leverages its proprietary platform to manufacture synthetic DNA, which can later be used in various applications and can be mostly broken down into 1) Syn Bio Write, 2) NGS Read, 3) Biopharma Solutions, and 4) DNA Data Storage. As of the latest quarterly data, these segments represented 37.5%, 55.2%, and 7.3%, respectively, as DNA Data Storage isn't generating revenues yet.

The company's revenue-generating segments are promising already. Still, I believe TWST's DNA Synthesis Platform might have the highest potential yet, as it uses DNA synthesis to overcome DNA obtainment limitations. This effectively escalates DNA production and reduces costs. Moreover, this platform's technology applies techniques from semiconductor production to improve efficiency at a lower price.



Source: 2024 AGBT Investor Dinner Presentation - February 2024.

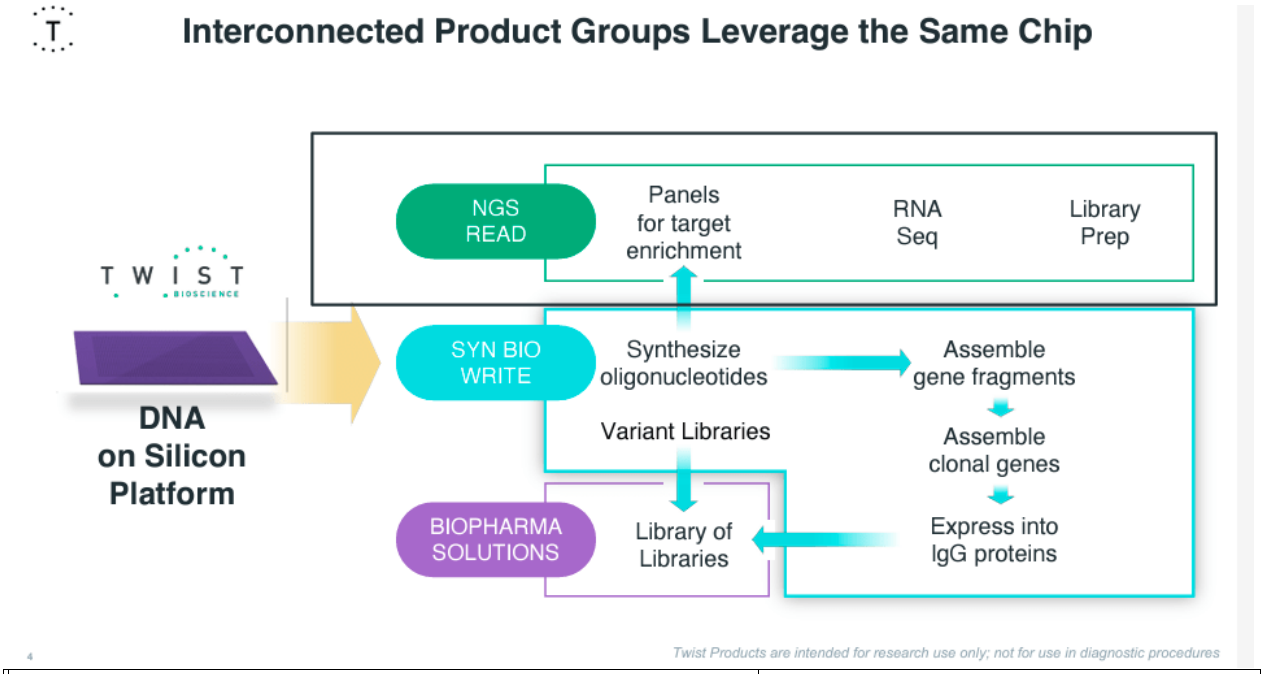
TWST's platform uses silicon to create arrays of synthetic DNA to process up to millions of DNA sequences in parallel with high fidelity. As a result of automating this process, artificial DNA has become accessible, fulfilling the needs of researchers who require precise, large-scale DNA production. Such DNA production is key for various applications like therapeutics and drug discovery, test diagnostics, bio-based chemicals, engineering crops, and, as previously mentioned, potentially even digital storage for large amounts of data. Essentially, TWST's products can be used as tools for researchers in several biotechnology fields to manipulate genetic material. I believe it's a game-changing technology still in its early stages.

Furthermore, TWST applies its technology to deliver a range of developments, including synthetic genes designed according to specifications, clone genes that are copies of precise DNA sequences, gene fragments for portions of DNA to be used for applications like gene or enzyme construction, and oligo pools for CRISPR gene editing, or Next-Generation Sequencing libraries. The company also has a service called Express Genes for the fast delivery of artificial genes. The NGS tools and systems for preparing library kits and sequencing panels for fast DNA and RNA synthesis are also important contributions. These applications further demonstrate the flexibility and wide potential that TWST's IP has, which I think should reasonably add optionality to its valuation.

Recent Performance and Strategic Expansion

It's also worth noting that TWST's Q1 2024 report once again came in above guidance, which is usually a positive sign for a company. The latest quarterly filing showed revenue growth and hit a new record of $71.5 million. TWST's orders rose to $77 million, increasing its customer base, mostly attributed to the ongoing trend of clients adopting TWST's new technologies in segments like synthetic genes, oligo pools, DNA libraries, and NGS.

Additionally, these revenues are highly profitable, with over 40% gross margins, which shows pricing power. Still, it's worth noting that TWST is not yet profitable, but its EBIT margins have consistently improved. I noticed that TWST's operating expenses are relatively stable, around $50 million for the last few quarters, and haven't increased alongside its growing revenues. This is why EBIT margins are continually improving QoQ, and I think there's a reasonable path towards profitability as long as this trend continues.

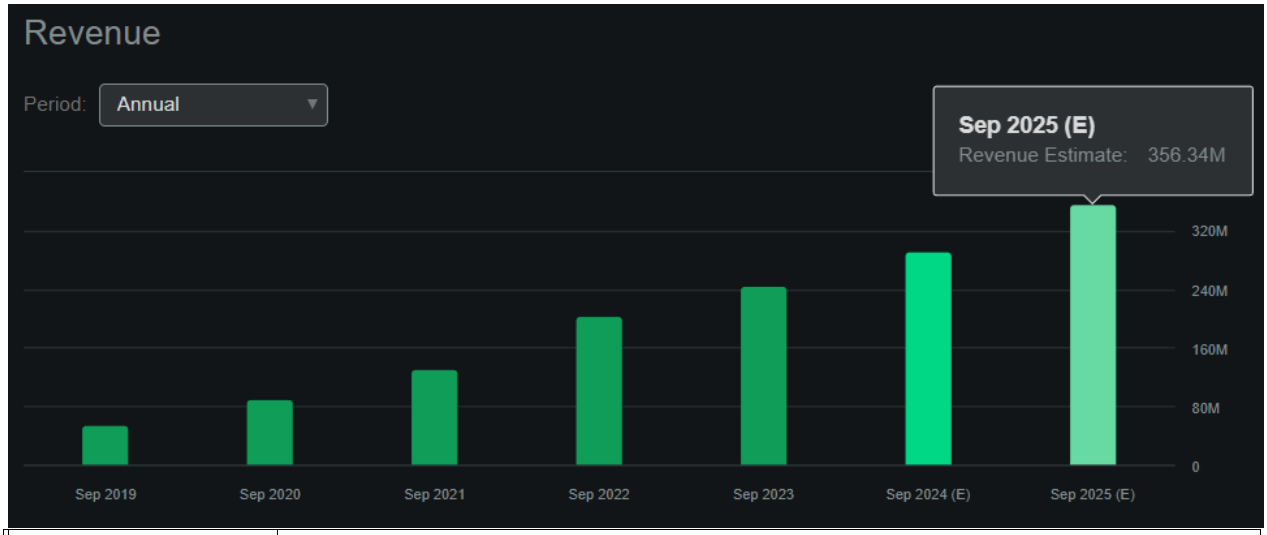


Source: 2024 AGBT Investor Dinner Presentation - February 2024.

Lastly, the company's Express Genes service expanded, capturing more researchers as customers. This was mostly because of TWST's five-day shipping turnaround for larger DNA preparations of up to 1 milligram. Interestingly, these developments occurred while slightly curtailing R&D expenses, tilting the company's efforts towards commercializing its IP.

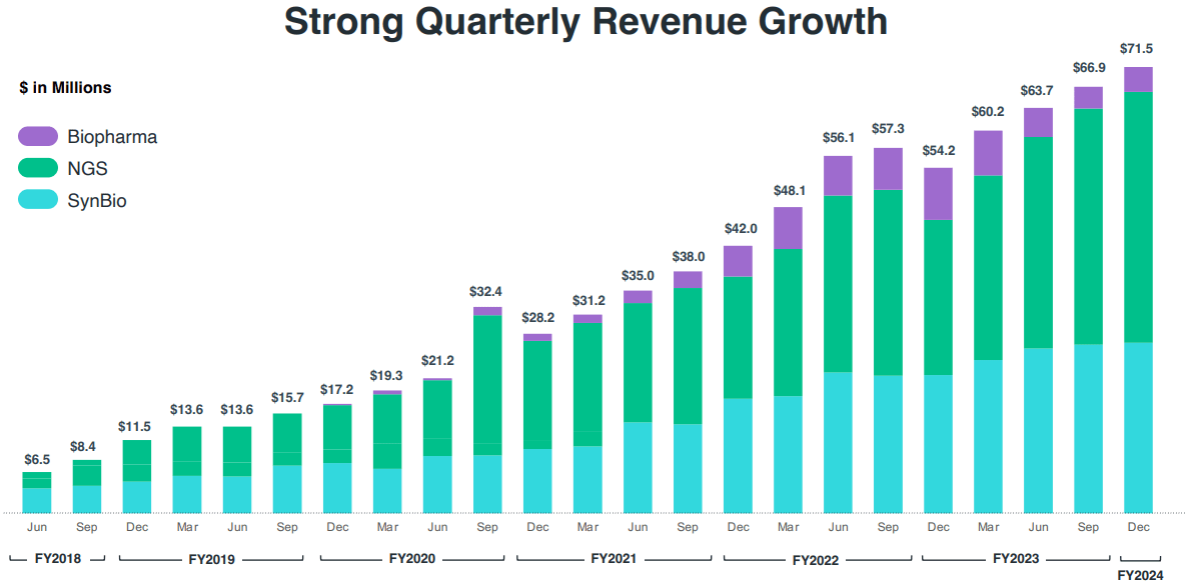
Justifiably Expensive: Valuation Analysis

From a valuation perspective, TWST is not a microcap biotech worth a few hundred million dollars in market cap hoping for FDA approval. TWST is slightly more mature than that, with consistently growing revenues and a respectable market cap of about $2.0 billion. However, if we annualize the latest quarterly revenues, TWST's current revenue run rate is approximately $286.0 million annually. That would imply a high P/S ratio 7.06, notably higher than the sector's median P/S multiple of 2.06. Even Seeking Alpha's dashboard revenue forecasts for TWST in 2025 imply a valuation premium. By 2025, the company is projected to generate about $356.34 million in yearly revenues, which means a forward P/S multiple of 5.67. This multiple is still considerably higher than its peers, so TWST trades at a premium.



Source: Seeking Alpha.

Nevertheless, TWST's liquidity seems safe, with $311.1 million in cash, cash equivalents, and short-term investments. Furthermore, TWST's CEO expressed the company's focus on revenue growth and financial discipline to achieve profitability and sustainable success. So, I think TWST is currently pivoting into a monetization phase, caring about its financial sustainability over the long run. This is promising, and if we annualize its most recent quarterly cash burn rate, TWST does appear healthy for now. I estimate TWST's quarterly cash burn to be about $24.5 million by adding up its cash flow from operations and CAPEX. This would mean TWST burns through $98.0 million annually, implying a cash runway of roughly 3.2 years. However, since TWST's revenues should continue growing and its focus is on cost management at this stage, I'd argue that this is a conservative estimate.



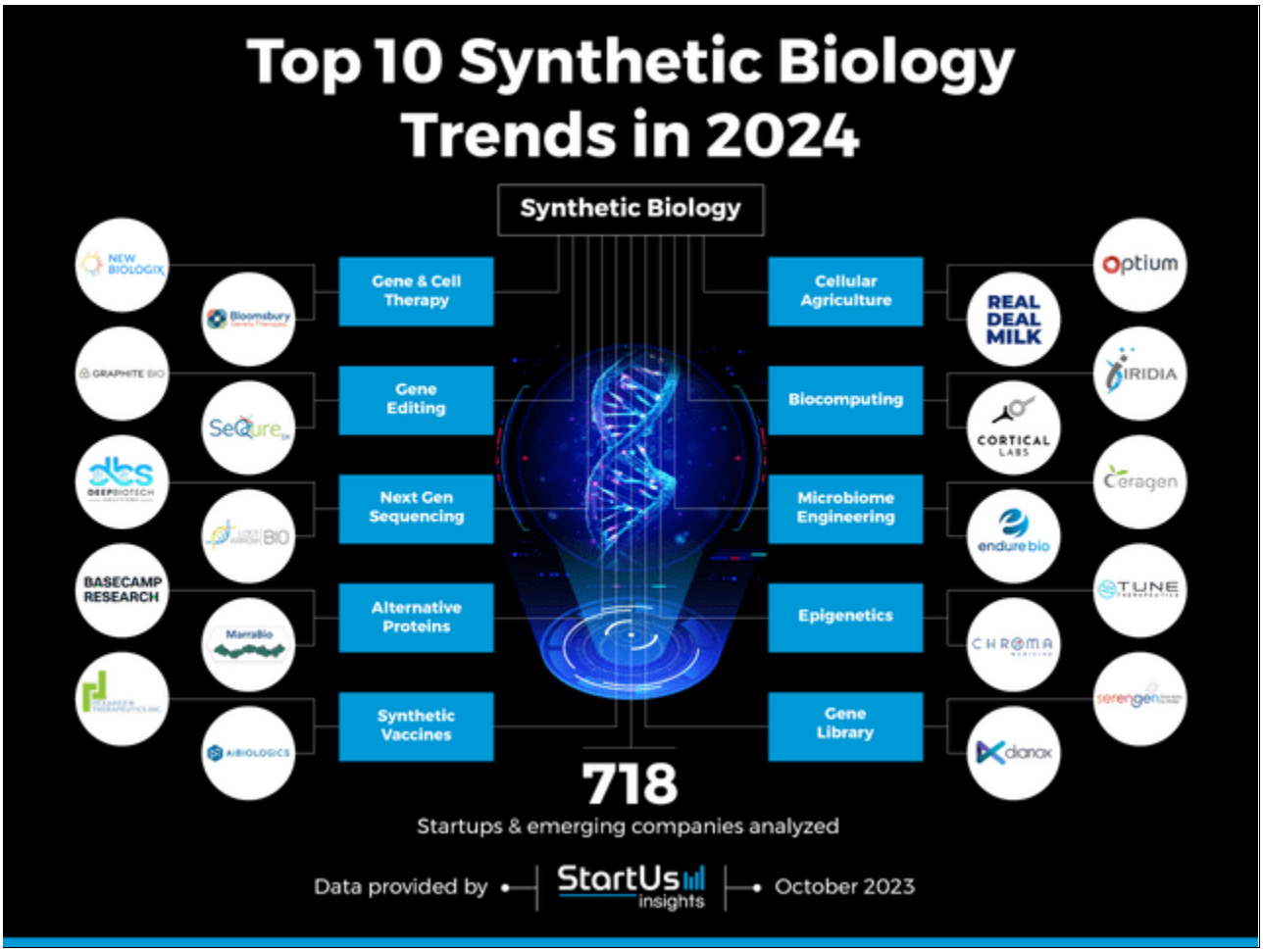
Source: TWST's Fiscal 2024 1Q Financial Results.

As you can see in the image above, the company has had an impressive growth rate since 2018. Using the figures from that chart, I estimated a CAGR of about 54.7% since June 2018. This is, without a doubt, explosive growth and shows that TWST undoubtedly deserves a growth premium. This is particularly true after you consider that DNA data storage isn't even being monetized at this stage, and that could potentially be a game changer down the road. After all, DNA storage could be much more cost-efficient for certain types of data that don't require frequent read-and-write operations. For instance, long-term data storage would greatly benefit from DNA's denser and more durable properties than traditional cloud storage.

So, putting it all together, TWST is a promising biotech company with impressive top-line growth that trades at a relatively elevated valuation multiple. This makes me lean towards a more neutral assessment of the company, as I'm incredibly bullish on the underlying tech and business fundamentals. Yet, simultaneously, my optimism is tempered by the already inflated valuation. This leads me to a "hold" rating, but I lean on the bullish side. But for now, I think this is a "buy on dips" stock rather than a "great buy" at this time.

TWST's Risks: Company and Macro Perspective

Moreover, I've identified a few caveats that further detract from the "buy" rating on TWST. First, I believe that the recent efforts to cut costs and focus on a path toward profitability are commendable from a financial perspective but may risk the company's long-term prospects. Synthetic biology and DNA applications remain a rapidly evolving field, and without constant R&D expenditures, there's a real risk that TWST's IP portfolio might become outdated relatively quickly. Also, while DNA applications are promising, other competing narratives might attract more investor capital if TWST's tech becomes stale or meaningful progress halts. Hence, this company might not be ideally suited to look for profitability just yet. Instead, it may be better to focus on finding a patent to achieve a lasting business moat.



Potential competing narratives. (Source: StartUs Insights.)

However, today, I concede that TWST's platform, particularly regarding DNA synthesis, is almost without a match. TWST leverages this to offer a wide range of applications, from healthcare to agriculture, at lower costs and higher speeds, making it more scalable. Given that the company's business is centered around synthetic biology, we estimate its current market size to be around $11.4 billion (Markets and Markets, 2022). Moreover, the source also reported a sector-wide CAGR of 25.6% until 2027, below TWST's realized CAGR since 2018. This could imply that TWST's growth rate will likely decline, trending towards the sector's CAGR. However, the flip side of that argument is that TWST's above-average CAGR corroborates its strong value proposition with its customers. We don't know at this time, so it's a risk that new investors would take on if they invest at these levels.

Conclusion: A Great Buy On Dips

Overall, TWST is one of the most promising biotech stocks I've researched. It's monetizing its platform and has consistently grown revenues at an impressive CAGR for several years. Moreover, I think there's still much untapped potential in its IP, especially in DNA Data Storage. This alternative data storage method could become a game

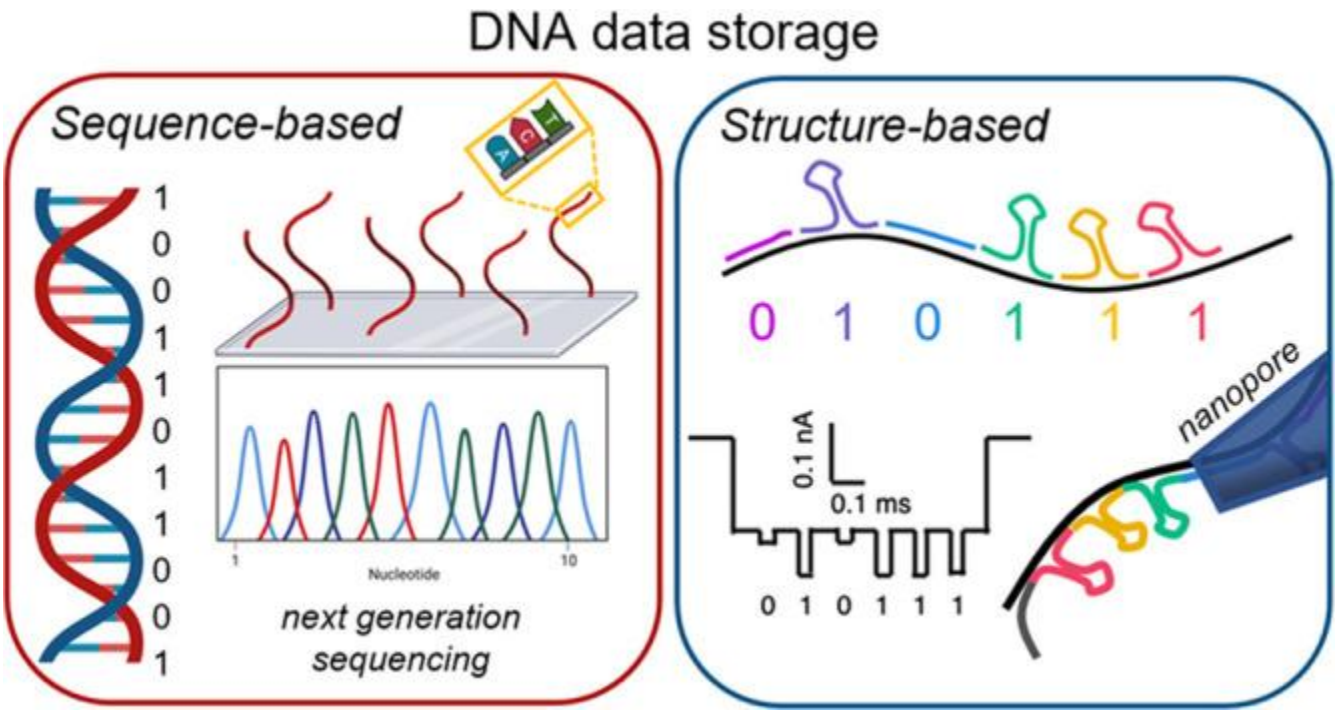
changer for certain storage needs. Also, TWST's balance sheet seems solid, and its cash runway also appears safe. Yet, at the same time, it's undeniable that the stock is already trading at a high valuation multiple. This tempers my enthusiasm for the stock, not the business itself or its technology. Hence, I rate TWST a "hold," but I still consider it a great buy on dips.

Emerging Approaches to DNA Data Storage: Challenges and Prospects

Andrea Doricchi [†,‡], Casey M Platnich [§], Andreas Gimpel [‖], Friederikee Horn [⊥], Max Earle [§], German Lanzavecchia [†,#], Aitziber L Cortajarena [■,○], Luis M Liz-Marzán [■,○,△], Na Liu [▼,⬡], Reinhard Heckel [⊥], Robert N Grass [‖], Roman Krahne [†], Ulrich F Keyser [§,*], Denis Garoli [†,*]

PMCID: PMC9706676 PMID: 36256971

Abstract



With the total amount of worldwide data skyrocketing, the global data storage demand is predicted to grow to 1.75×10^{14} GB by 2025. Traditional storage methods have difficulties keeping pace given that current storage media have a maximum density of 10^3 GB/mm^3. As such, data production will far exceed the capacity of currently available storage methods. The costs of maintaining and transferring data, as well as the limited lifespans and significant data losses associated with current technologies also demand advanced solutions for information storage. Nature offers a powerful alternative through

the storage of information that defines living organisms in unique orders of four bases (A, T, C, G) located in molecules called deoxyribonucleic acid (DNA). DNA molecules as information carriers have many advantages over traditional storage media. Their high storage density, potentially low maintenance cost, ease of synthesis, and chemical modification make them an ideal alternative for information storage. To this end, rapid progress has been made over the past decade by exploiting user-defined DNA materials to encode information. In this review, we discuss the most recent advances of DNA-based data storage with a major focus on the challenges that remain in this promising field, including the current intrinsic low speed in data writing and reading and the high cost per byte stored. Alternatively, data storage relying on DNA nanostructures (as opposed to DNA sequence) as well as on other combinations of nanomaterials and biomolecules are proposed with promising technological and economic advantages. In summarizing the advances that have been made and underlining the challenges that remain, we provide a roadmap for the ongoing research in this rapidly growing field, which will enable the development of technological solutions to the global demand for superior storage methodologies.
Keywords: DNA, data storage, sequencing, random access, error correction, DNA nanostructure, DNA preservation, reading, decoding, costs

1. Introduction

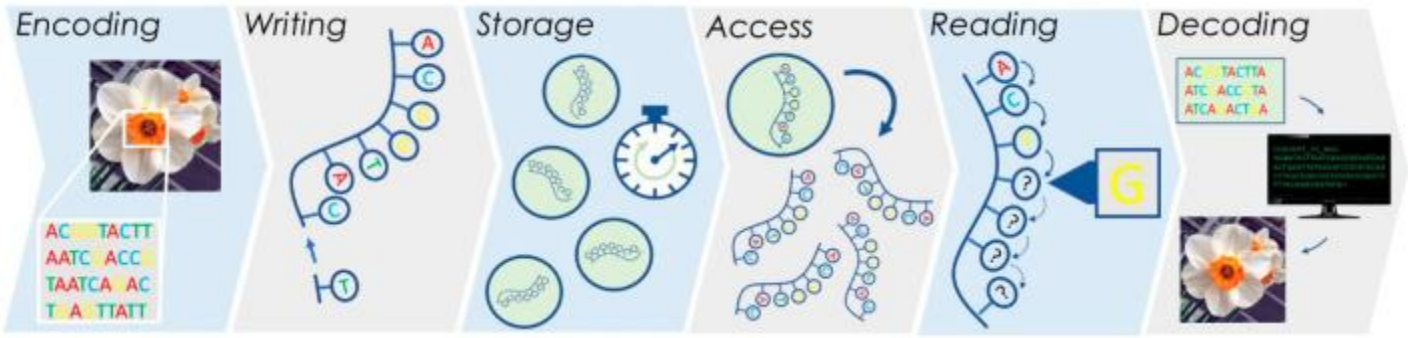
In the present digital era, the quantity of data being produced continues to increase exponentially, with the global demand for data storage expected to grow up to 1.75 × 10^{14} GB by 2025 and by a further order of magnitude within the end of this decade.[1] The demand for denser and longer-lived information storage devices is also increasing.[2] Current storage technologies, including optical and magnetic devices, are reaching their information density limits and are thus not suitable for long-term (>50 years) storage, which means that valuable information needs to regularly be transferred to newer storage media if it is to be preserved for future generations. Innovative methods are required for long-term information storage to circumvent this laborious and costly process and to combat other pitfalls associated with current storage media (including energy consumption and insufficient data density).[3]

Nature provides an inspiring example of how to encode, transmit, and preserve information by using DNA to store all genetic information in the form of a four nucleotide sequence. As evidenced by DNA's invaluable role in the perpetuation of genetic information, these molecules are stable for thousands of years under suitable storage conditions;[4] for example, 300 000-year-old mitochondrial DNA from a bear has been successfully sequenced.[5] This DNA sample was preserved in bone, thereby demonstrating that the required power consumption for the archival storage of DNA is very low—another benefit compared with traditional data storage media. In addition to its stability and low cost of storage, DNA presents a major key advantage compared with existing data storage devices: data density. On the basis of its physical dimensions, DNA has a theoretical data density of 6 bits for every 1 nm of polymer, or ~4.5 ×

10^7 GB/g,[6] which is orders of magnitude higher than the densities achievable using traditional devices.[7,8]

Significant advances have been made in recent years toward using DNA as a digital information storage medium.[9−14] Existing strategies to encode arbitrary information into DNA do so by translating the desired data (i.e., a movie, book, or picture) directly into the nucleotide sequence, which means that to write each data string, chemical DNA synthesis is employed.[15] In sequence-based DNA data storage, the major steps comprise: (1) encoding digital information, (2) data writing (synthesis of new oligonucleotides), (3) storing the DNA in physical or biological conditions, (4) random access, (5) data readout via DNA sequencing, and (6) decoding the DNA sequences back into the original digital code, as represented in Figure 1.[8,12] Over the past decades, substantial advances in biotechnology have significantly bolstered DNA data storage technologies. These include chemical and enzymatic DNA syntheses,[16,17] polymerase chain reaction (PCR) for DNA amplification,[18] and DNA sequencing.[19] Although none of these technologies was initially designed with digital data storage in mind, these considerable developments now enable procedures for writing, random accessing, reading, and editing of data encoded in DNA sequences.[10,11]

Figure 1.

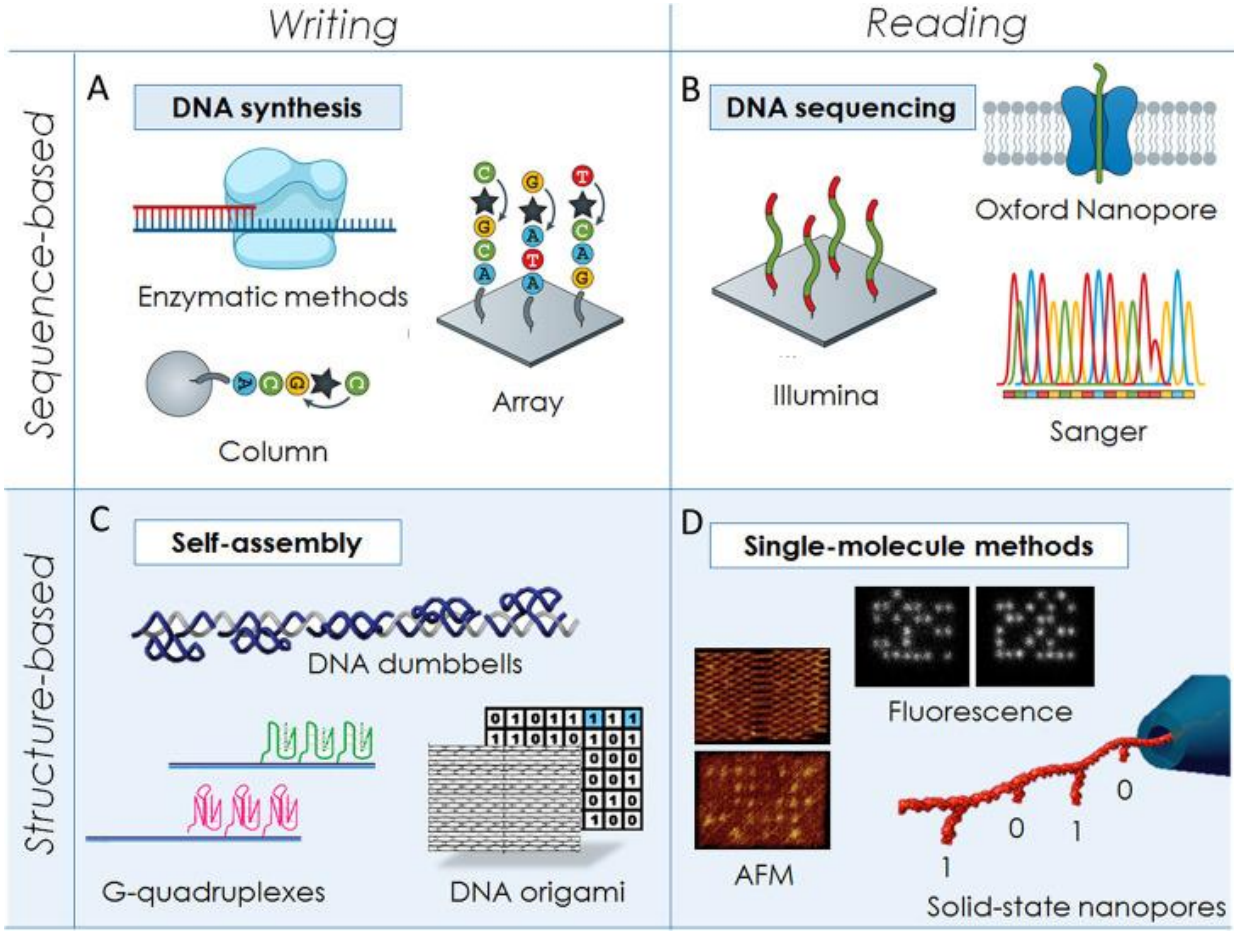


General strategy for DNA data storage, wherein the data is stored directly in the sequence of the oligonucleotides. The six main steps—encoding, writing, storage, access, reading, and decoding—are depicted.

However, each of the procedures involved in DNA data storage—encoding, writing, storage, random access, reading, and decoding—has significant technical limitations that render DNA data storage, at present, not competitive with magnetic and solid-state storage devices. Because the *de novo* synthesis of long sequences of DNA remains challenging,[20] these sequences must be broken into smaller fragments (~200 bases), which requires massive numbers of unique DNA sequences to be made. Data readout also presents several challenges: while in theory analogous to the magnetic readout of a hard disk drive, DNA sequencing must be employed to read out the information stored in individual oligonucleotides. Sequencing often relies on fluorescence outputs, which require expensive fluorophores, optical equipment, and trained personnel, as well as substantial amounts of DNA and long reading times (Figure 2a,b). Nanopore methods

may present an appealing alternative, as detailed further in this review. With the use of current technologies, DNA storage is estimated to cost 800 million USD per one terabyte of data (by contrast, tape storage costs approximately 15 USD per terabyte).[12] The high price of writing DNA data using existing methods prohibits its mainstream adoption as an information storage material.

Figure 2.



Comparison of the main differences between sequence-based (A,B) and structure-based DNA data storage (C,D), as has been presented in the literature to date. (A,B) Sequence-based storage relies on the *de novo* synthesis of DNA strands and the subsequent sequencing of these entities is performed using next-generation methods. Image adapted with permission from ref (12). Copyright 2019 Springer Nature. (C) By contrast, structure-based methods utilize self-assembly, which means that the information is encoded into their three-dimensional shape. Images adapted with permission: ref (21), copyright 2016 Springer Nature; ref (22), under a Creative Commons Attribution 4.0 License (CC BY), copyright 2021 Springer Nature. (D) These shapes can then be read off using single-molecule methods, including fluorescence, atomic force microscopy, and nanopore techniques. Image adapted from ref (23). Copyright 2019 American Chemical Society.

One potential strategy to circumvent these pitfalls is to rely on the programmable three-dimensional structure of DNA as opposed to its primary sequence (Figure 2c,d). DNA nanotechnology harnesses the specific base-pairing properties of the nitrogenous bases to create arbitrary two- and three-dimensional shapes.[24] It is possible to generate well-defined, custom objects at the nanoscale using these methods. Information can thus be stored in the 3D structures of these assemblies instead of in the sequence, with readout relying on imaging techniques, such as super resolution imaging,[22] or using single-molecule nanopore measurements.[23,25] The structure-based strategy may reduce the number of DNA sequences that must be synthesized by allowing for the erasing and rewriting of data through simple self-assembly. These structure-based methods also eliminate the need for next-generation sequencing, which remains among the most time-consuming aspects of DNA data storage. Because DNA nanotechnology-based approaches capitalize on the self-assembly of DNA sequences, the resulting structures are inherently reconfigurable, which enables data erasing and rewriting without further synthesis.[26] Moreover, the dynamic nature of these assemblies can be exploited to perform data operations,[13,27] which allows DNA data storage to integrate directly into the field of DNA computation.

In this review, we provide a detailed description of the two aforementioned methods, which we will refer to as “sequence-based” and “structure-based” DNA data storage. A comparison between them that highlights both the similarities and differences in these approaches will provide an overview of the state of the art in DNA data storage. Finally, we also highlight the exciting potential applications of DNA data storage and manipulation, including archival storage, barcoding, cryptography,[11] and DNA computing. Despite the hurdles that must be surmounted to implement DNA data storage, it is important to remember that DNA plays an irreplaceable role in biological systems. As such, DNA will never become obsolete as a data storage medium. We posit that the fundamental nature of DNA, in combination with the high density and low energy cost of DNA data storage, will continue to fuel research in this rapidly growing domain.

2. Sequence-Based DNA Data Storage Methods

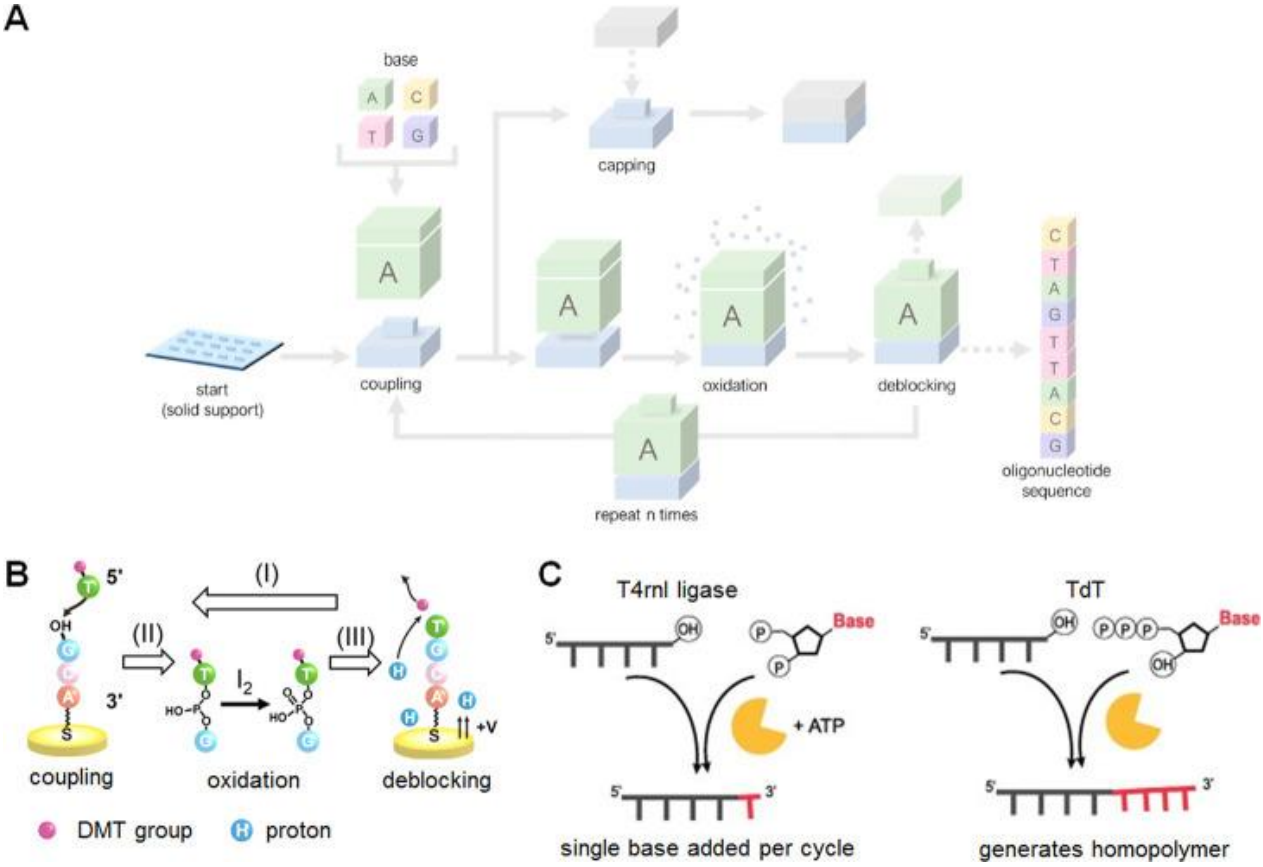
2.1. From Encoding to Data Writing in DNA Data Storage

Any digital data (files of any kind such as text and pictures) can be represented as a sequence of bits (i.e., zeros and ones). One possible data storage approach is to use a set of DNA sequences of 60–200 nt in length. The limitations in sequence length arise from the chemical synthesis of DNA; producing DNA strands longer than a few hundred nucleotides (nt) introduces a significant number of errors into the sequence.

Once properly encoded, data are written on synthetic DNA sequences (Figure 3). Organic chemistry has presented us with a large set of techniques for synthesizing DNA and, as previously mentioned, strands up to 200 nt in length can be readily synthesized. The synthesis is typically performed using phosphoramidite chemistry, which is a four-

step cyclic reaction involving the addition of the desired nucleotide to a growing oligonucleotide chain immobilized on a solid support (Figure 3A,B).[28] The use of a solid support enables extensive parallel synthesis, as well as automation of the chemical process, which will be fundamental to the adoption of DNA for data storage applications.[29,30] While there are many advantages to phosphoramidite synthesis, it is worth noting that it requires the use of anhydrous solvents, which produce toxic waste.

Figure 3.



An overview of chemical and enzymatic strategies to synthesize custom DNA sequences. (A) Phosphoramidite synthesis—the most widely used chemical strategy for the synthesis of DNA—involves the sequential addition of nucleotides to a growing chain anchored on a solid support. Protecting groups are employed to ensure that no more than one nucleotide is added at each step and are then subsequently removed via chemical deblocking. (B) Deblocking can also be performed by electrochemistry. Reproduced with permission under a Creative Commons Attribution 4.0 License (CC BY-NC) from ref (31). Copyright 2021 AAAS. (C) Enzymatic methods relying on T4rnl ligase or TdT can also be used to specifically add bases to a growing oligonucleotide in aqueous environments, which eliminates the need for organic solvents. Image reproduced with permission under a Creative Commons Attribution 4.0 License (CC BY) from ref (32). Copyright 2021 Elsevier B.V.

An alternative to chemical synthesis is enzyme-based methods, but they are still in their infancy. So far, only tiny amounts of data (hundreds of bits) have been stored using enzymatic synthesis versus data consisting of billions of bits using phosphoramidite synthesis. The concept of enzymatic DNA synthesis arose from the discovery of specific DNA polymerases, and this approach is expected to become both cheaper and faster than phosphoramidite synthesis for data storage applications.[40] A major limitation, however, is DNA polymerase's need for a template strand. To create a user-defined DNA sequence as in the chemical method, enzymes capable of extending the 3′ end of the ssDNA in a template-independent manner, such as polynucleotide phosphorylase (PNPase), T4 RNA ligase, and terminal deoxynucleotidyl transferase (TdT, Figure 3D), are required.[32] In particular, the use of TdT, a template-independent polymerase, to synthesize DNA oligonucleotides was shown to be a promising alternative to chemical synthesis.[33,16] Among others, Lee et al.[16] reported on a technique for enzymatic synthesis and digital coding that was based on template-independent polymerase TdT and nanopore reading. This strategy allowed the archiving of information in DNA without mandatory single-base precision, as well as cost reduction due to miniaturization and enzyme recycling. Moreover, the synthesis of 1000-nucleotide-long strands with homopolymeric stretches enabled a reduction of the synthesis time (Figure 3D). Palluk et al.[28,33] also described an oligonucleotide synthesis strategy that uses TdT and demonstrated that TdT–dNTP conjugates can quantitatively extend a primer by a single nt in 10–20 s. Crucially, this scheme can be iterated to write a user-defined sequence. Compared with chemical synthesis, which is undertaken in organic solvents, the enzymatic synthesis is compatible with aqueous conditions.

Both chemical and enzymatic syntheses are severely limited by the low speed of these processes.[29,39] Achievement of the necessary parallel writing capabilities while maintaining a realistic infrastructure footprint requires maximization of the number of different sequences that can be synthesized per unit area, simultaneously, on a single platform. The most space-efficient way to increase synthesis density is to reduce the area over which each unique sequence is grown (the feature size), the distance between features (the pitch), or both. To this end, photomask arrays have proven to generate high oligonucleotide densities;[34] however, this technique relies on a series of bespoke photolithographic masks to synthesize a defined set of sequences, that is, masks must be created for each set of desired sequences. An alternative method uses electrode arrays and leverages the scaling and production roadmap of the semiconductor industry, where features as small as 5 nm are now common. For example, Nguyen et al.[35] produced an electrode array and demonstrated independent electrode-specific control of DNA synthesis with electrode sizes and pitches that enabled a synthesis density of 25 million oligonucleotides/cm^2 (Figure 3C). Finally, the printing synthesis method has rapidly become the most applied method (also thanks to commercial technological platforms, such as Agilent and Twist).

The sequences to be synthesized are defined by the encoding process, which maps the data to a set of DNA sequences so that a corresponding decoder can reconstruct the

information, even though the writing, reading, and storage of the DNA introduces errors.[9,14,36,7,37,29,38,16,39−41]

DNA storage systems overcome these errors without losing data by capitalizing on both physical and logical redundancy. Physical redundancy is achieved by creating many, sometimes inaccurate, copies of each sequence, which enables a consensus to be reached when the data is read. Some errors cannot be resolved using physical redundancy alone. Logical redundancy guarantees reconstruction even when errors occur. While physical redundancy occurs automatically during the synthesis process—many copies of each sequence are always produced—it is fundamental to apply dedicated algorithms to include logical redundancies in the initial encoding. Moreover, encoding and decoding are strictly connected. The algorithms that encode the data to be stored add redundancy in a principled way so that a decoding algorithm can reconstruct the data from noisy reads (Figure 3A).

During the 2010s, extensive innovations in algorithm development have enabled reliable storage of data even under significant errors. Grass et al.[4] used modern error-correcting codes in the context of DNA storage, and a variety of different schemes have been proposed.[4,42−50,36,14] While physical and logical redundancy lower the storage density of DNA, recent works have proposed to raise it by expanding the DNA alphabet using composite natural letters[7,51,52] or chemically modified nucleotides.[53]

2.2. Storage and Degradation Issues

Despite DNA's long-term stability in well-controlled environments such as ancient bone, with storage durations as long as several hundred thousand years,[54,55] both aqueous solutions and dried DNA only exhibit a half-life on the order of months to a few years under ambient conditions.[56] Therefore, considerations for the physical storage of data-encoding DNA are crucial for realizing its potential for long-term data storage. Without appropriate protection, DNA (and thus the data encoded within) is degraded by multiple mechanisms, including strand breaks, nucleotide mutations, strand cross-linking by UV, oxidation, hydrolysis, alkylation, or mechanical stress, all of which are due to environmental factors. Among those, hydrolysis is the dominating decay pathway in a data storage context.[57,58] Thus, all applicable DNA storage approaches focus on protecting the DNA from moisture and oxygen with either microscopic (i.e., on the level of individual molecules) or macroscopic (i.e., on the level of individual pools) containers. Examples of microscopic containers include encapsulation within silica particles;[56,59−61] embedding in alkaline salt,[62] polymer,[63] sugar,[64] or silk protein[65] matrices; and coprecipitation with calcium phosphates[66] imitating bone. In the latter category, dried or lyophilized DNA is stored on filter paper[64] within hermetically sealed capsules with inert atmosphere[57,58,67,68] or, as is common in biological practice, simply frozen in aqueous solutions and stored at −20 or −80 °C.[69]

Generally, all storage approaches trade long-term stability with a decrease in storage density by 1–3 orders of magnitude, caused by the low loading ratio between DNA and carrier (see Table 1). Additionally, the required time and cost for protection can be a distinguishing factor for DNA data storage systems, albeit less so for long-term storage applications.[69] The size of a single DNA pool is an important consideration for the design of DNA storage media, as index sizes for random access, constraints of PCR, and required physical redundancy for retrieval imply an upper limit on the number of pooled oligos.[69,6] This represents the maximum data that can be stored within a single macroscopic storage container, and has been estimated to lie between a few TB up to a few hundred TB.[56,6,70] We compared the storage densities and half-lives of micro- and macroscopic storage approaches in Table 1 by using the largest model pool size for which random access has been demonstrated at 5.5 TB.[71]

Table 1. Comparison of the DNA loadings (g DNA/g carrier), achieved information density in PB/g, and extrapolated half lives for both macroscopic and microscopic storage approaches, with an assumed pool size of 5.5 TB.[71,a].

storage approach	τ(10 °C)/ years	DNA loading[c]	density[c]/PB/g	references
		macroscopic		
in solution[b]	17	0.005%	0.85	(57)
dried	7	100%	17 000	(66)
bone	1700	0.05%	8.5	(54,62)
DNAshell[c]	>100 000	0.000 02%	0.0034	(58,68)
		microscopic		
trehalose matrix	160	0.13%	2.2	(64)
silica particles	540	3.4%	580	(59,4)
polymer matrix[d]	110	0.1%	17	(63)
salt matrices	750	20%	3400	(62)
silk matrix[e]	NA	0.000 03%	0.0051	(65)
calcium phosphate matrix	600	18%	3060	(66)

a

All values are considered at 10 °C and assuming DNA with 150 bp at an information density of 17 EB/g. Temperature corrections were performed using Arrhenius Law using 155 kJ/mol as the activation energy of DNA strand breaks.[57,4]

b

Typical concentration for synthetic DNA is 500 ng/μL.
[c]

Assumed pool size per DNA shell = 5.5 TB.[71] Weight of the DNA shell is at least 1.3 g.[58]
[d]

Polymer density was assumed as similar to that of polyethylene glycol at 1.12 g/cm^3.
[e]

Density of filter paper is around 85 kg/m^2.

Current approaches towards data encoding in DNA, such as the use of altered DNA topology[23,72] and third-generation sequencing platforms, present new challenges to data storage, as those approaches rely on oligos with multiple hundreds to thousands of nucleotides in length, compared with the few hundreds of nucleotides commonly in use for next-generation sequencing (NGS).[70] While both micro- and macroscopic storage systems are independent of sequence length, DNA decay by hydrolysis scales with the number of nucleotides per oligo and, thus, a proportional increase in the expected number of errors is anticipated.[73] Given that some types of single-site errors, such as strand breaks, may render entire oligos and the data within unreadable, the use of longer sequences further increases the need for durable storage to prevent premature data decay beyond experimental time scales. To this end, systematic studies on decay mechanisms and rates for many approaches to data encoding in DNA are missing, a critical factor regarding approaches that heavily rely on structural integrity for data retrieval.

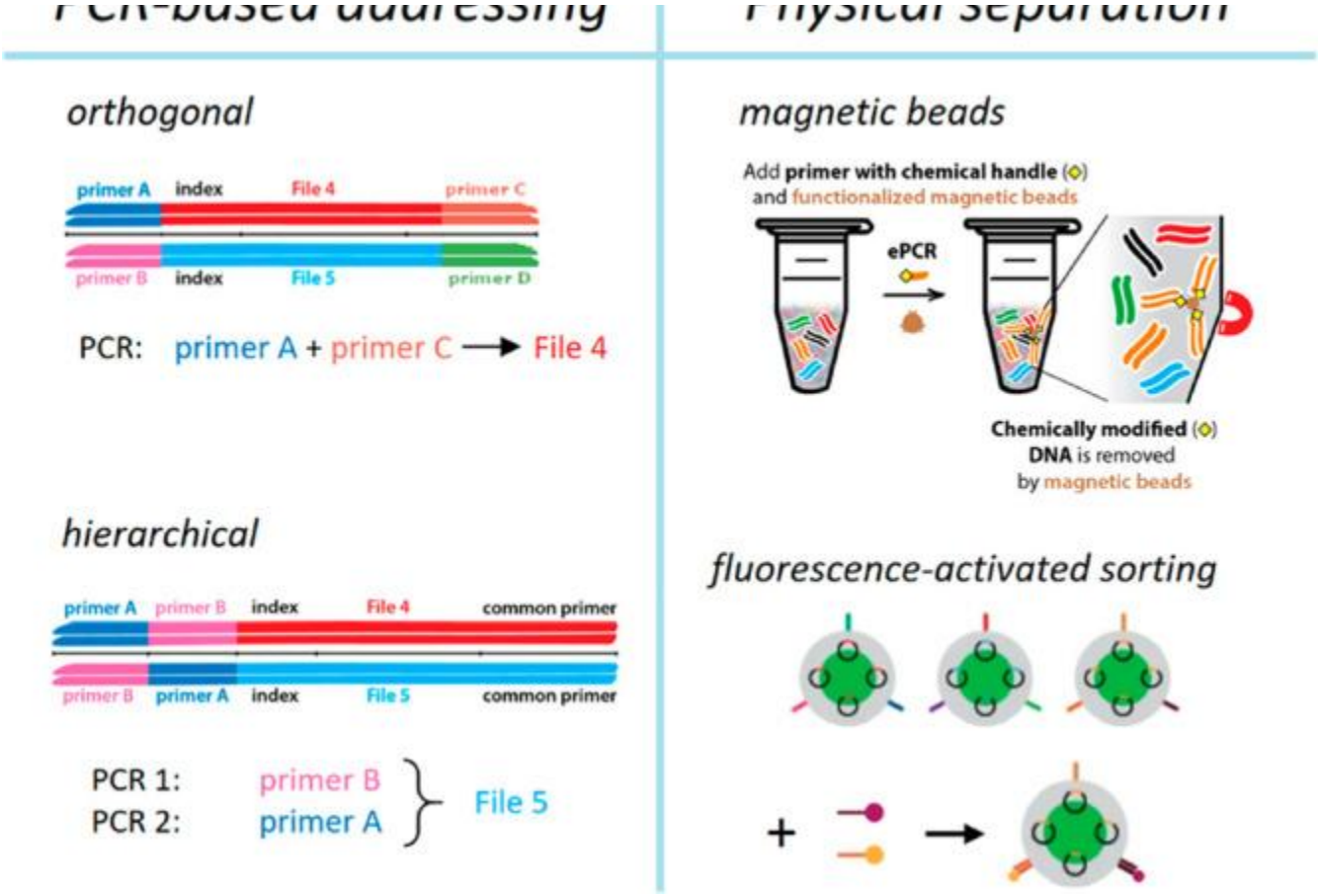
Currently, long-term storage is only feasible within a protective material and at DNA loadings of only a few percent. Consequently, the need remains for long-term DNA data storage systems closer to DNA's true storage density. Indeed, further improvements in the coding density toward DNA's Shannon capacity, for example, by means of improved encoding algorithms or lowering logical redundancy, are largely overshadowed by the general loss of storage density due to the storage matrix. Conversely, the loss in encoding density yielded by encoding approaches relying on DNA topology is rendered less severe by this storage overhead, and the interplay of such approaches with denser storage systems is interesting for further research.

2.3. Random Access

As discussed above, the ability to select only a subset of DNA molecules for readout limits the current data capacity of a single pool of data-encoding DNA. This access to DNA subpools, equivalent to file-level random access, is crucial to scale DNA-based data storage up to large data capacities with no need for costly, complete sequencing of the pool. This has a major implication: an addressing system is needed to select

subpools from a complex DNA mixture, with high specificity. Whereas the use of a physical substrate on which DNA can be arrayed may solve this problem,[31] this approach and similar solutions relying on the physical separation of individual oligos or oligo pools render DNA's density advantage obsolete. Instead, two other major strategies have been developed: PCR-based addressing and direct physical separation (Figure 4). In PCR-based addressing systems, the high specificity of amplification via PCR is leveraged to selectively enrich a subpool over the background by using at least one address-specific primer and corresponding priming regions on the data-encoding oligos. Because of PCR's exponential nature, a sample of the amplified pool will contain mainly the desired file with its matching priming regions, as well as nonspecific sequences as background. Demonstrated in 2015,[74] this addressing system has now been shown to scale to well above 10^{10} unique sequences per reaction while only requiring about 10 copies per sequence, which is equivalent to a pool capacity on the order of terabytes.[6,70,71] Either a rigorous design of orthogonal primer sequences[6] or the use of hierarchical addressing systems would be needed to achieve the required high specificity at these scales.[71] Nonetheless, primer-based addressing systems face several constraints. First, the incorporation of random-access priming regions into each oligo decreases the available space for data-encoding bases, thereby also decreasing the storage density (currently by about 15% per address region).[71,75] Second, PCR-based random access irreversibly removes oligos from the pool, which necessitates potentially lossy reamplification of the entire pool after repeated data retrieval.[75,76] Moreover, as pool sizes and, thus, the number of sequences, become larger, the enrichment of a few copies against an ever-increasing background will at some point hit the limitations of PCR regarding processing volumes, required amplification cycles to obtain sufficient enrichment, and nonspecific amplification due to primer–payload similarity.[77,70] Indeed, data retrieval from a hierarchical addressing system of 5.5 TB required additional physical separation of pools via a biotin-based bead extraction between file accesses to fully remove the background carried over from PCR.[71] Lastly, PCR-based addressing is incompatible with common storage approaches, thus necessitating the removal and re-embedding of the encoding DNA into the storage matrix for each random-access operation.

Figure 4.



Overview of random access strategies to select a subpool of sequences, usually a file, from a large pool. PCR-based addressing methods leverage the high specificity of primers and the exponential amplification of PCR to enrich target sequences by using either a single or multiple PCR runs. Methods using physical separation as a tool to select sequences also rely on the high specificity of short primers or barcode sequences, but remove the desired sequences using magnetic bead extraction or fluorescence-activated sorting. Images adapted from ref (71) and reproduced with permission from ref (75). Copyright 2019 American Chemical Society and copyright 2021 Springer Nature, respectively.

As an alternative to PCR, sequence specificity has also been exploited to carry out physical separation of files in pools. As mentioned above, biotin-labeled primers can be used to address and extract specific files via streptavidin magnetic beads on the basis of file-specific random access regions in encoding oligos, similarly to PCR-based addressing.[71,78] This approach has two key advantages: the sample can be reused for subsequent retrievals and nonspecific binding and PCR-induced biases are circumvented.[78] Banal et al.[75] extended this concept to DNA pools encapsulated in silica particles by labeling their surface with DNA barcodes to facilitate random access via fluorescently labeled probes and fluorescence sorting. While this represents a scalable random access scheme compatible with long-term storage, it is likely that the DNA

barcodes on silica particles would decay much faster than the data-encoding DNA within so that random access ceases to function even if the data itself may still be intact.

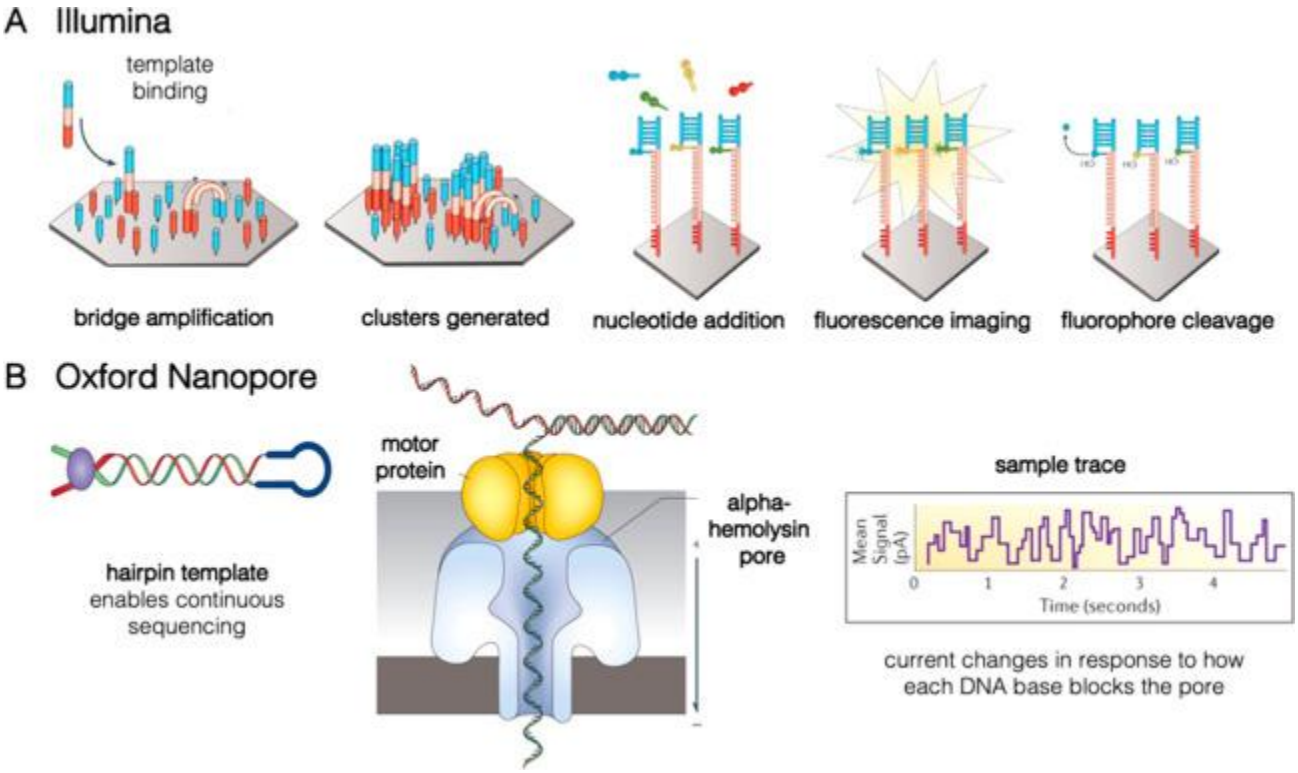
All random access approaches aim at facilitating file-level control in large pools of DNA-encoded files while under the constraints of specificity, scalability, and storage density. Such scaling to large pools is highly desirable because it retains DNA's high storage density compared with the physical separation of smaller pools using storage approaches (see Section 2.2). Currently, the highest demonstrated data capacity for random access is on the order of terabytes of data.[6,12,71] While this does not appear to be a hard limit,[6] it is likely unpractical to scale random access by PCR-based addressing indefinitely because of the aforementioned difficulty of orthogonal primer design and the requirement for many amplification cycles given the associated impact of PCR bias.[20] Whether any practical limit of PCR-based random access exists in real-life applications remains to be seen, however. As an alternative, hierarchical storage systems combining high-level access to isolated subpools with file-level random access within such subpools appear more suited to allow for random access at the data capacities envisioned for DNA data storage. The first steps in this direction have been taken, such as labeling DNA-embedding polymer disks with QR codes or automated retrieval of individual DNA pools in a digital microfluidic device,[56,63] but the trade-off between storage density, data longevity, and ease of automated data access requires further work.

Beyond random access, other file operations such as encryption with genomic keys,[79] erasure on the basis of obfuscation,[80] and rewriting by chemical modification or PCR[81,82] are also supported by sequence-based DNA data storage. As recently reviewed elsewhere,[83] these approaches highlight the versatility of file operations supported by DNA as a storage medium.

2.4. Reading

While the readout of data encoded in DNA is rarely done in its application as an archival storage system,[69] the complete and error-free retrieval of stored data must be guaranteed within a defined set of storage and sequencing conditions in order for DNA data storage to have any commercial relevance. As a result, the choice of sequencing platform has a marked impact on the design and feasibility of sequence-based DNA data storage. Currently, readout of the DNA sequences needed for data decoding relies heavily on established technologies for DNA sequencing in life science applications, most prominently sequencing-by-synthesis (SBS) as commercialized by Illumina.(Figure 5A,B).[84,85] As an alternative, sequencing using protein nanopores, commercialized by Oxford Nanopore Technologies, has been used because of its ease of implementation, automation, and portability (Figure 5C).[70,81,86] Nanopore sequencing uses electrical readouts rather than fluorescence detection to identify each base of a DNA strand as it moves through a biological nanopore. Contrary to SBS, it is therefore also able to identify modified and unnatural nucleotides such that the readout of data encoded using an expanded molecular alphabet is possible.[53,87]

Figure 5.



Overview of next-generation sequencing technologies presently used in DNA data storage. (A) Illumina sequencing generates clusters of identical single-stranded oligonucleotides. As the complement is synthesized using spectrally distinct, fluorescently tagged nucleotides, the identity of each base along the strand can be determined through the color of emission. (B) Oxford Nanopore measurements do not require fluorescent dye molecules. As the oligonucleotide passes through the protein pore, the three-dimensional shape of each base will modulate the ionic current, which results in a current–time trace that corresponds to the specific sequence. Images adapted with permission from ref (85). Copyright 2016 Springer Nature.

While nanopore sequencing improves upon several limitations of SBS for DNA data storage, as reviewed by Ceze et al.,[12] two key constraints of the technology are its high error rate and the required sequence length. The high error rate of nanopore sequencing (~10% per nt in the single read),[70,88] compared with the nearly negligible rate of errors introduced by SBS (~0.5% per nt),[70] necessitates the clustering of sequence information, and thus, higher sequencing coverage and additional postprocessing of sequencing data.[70,86] Moreover, sufficient pore utilization for high sequencing throughput can only be realized for long fragments (>1 kb).[86,88] Therefore, the readily available oligo libraries with a length of only a few hundred nucleotides per sequence must be combined into longer assemblies to be suitable for nanopore sequencing. This process, usually performed via Gibson assembly or overlap extension PCR,[70,81,86] reintroduces several difficult-to-automatize steps into the sequencing

workflow, which calls the approach's claims of improved portability and ease of automation over SBS into question. These constraints currently render nanopore sequencing more challenging and slower than SBS.[12] Accordingly, the largest data size retrieved using the technology is currently about 1.67 MB, compared with around 200 MB for SBS.[70,86]

The use of both state-of-the-art SBS and rapidly developing nanopore sequencing for DNA data storage highlights the current trade-off between sequencing accuracy and cost, as well as implications for future scalability. To this end, the development of solid-state nanopores for the determination of DNA structures including their sequence, with the potential of increased accuracy and throughput by avoiding enzymes limiting the translocation rate, holds promise for data storage applications.

2.5. Decoding and Error Correction

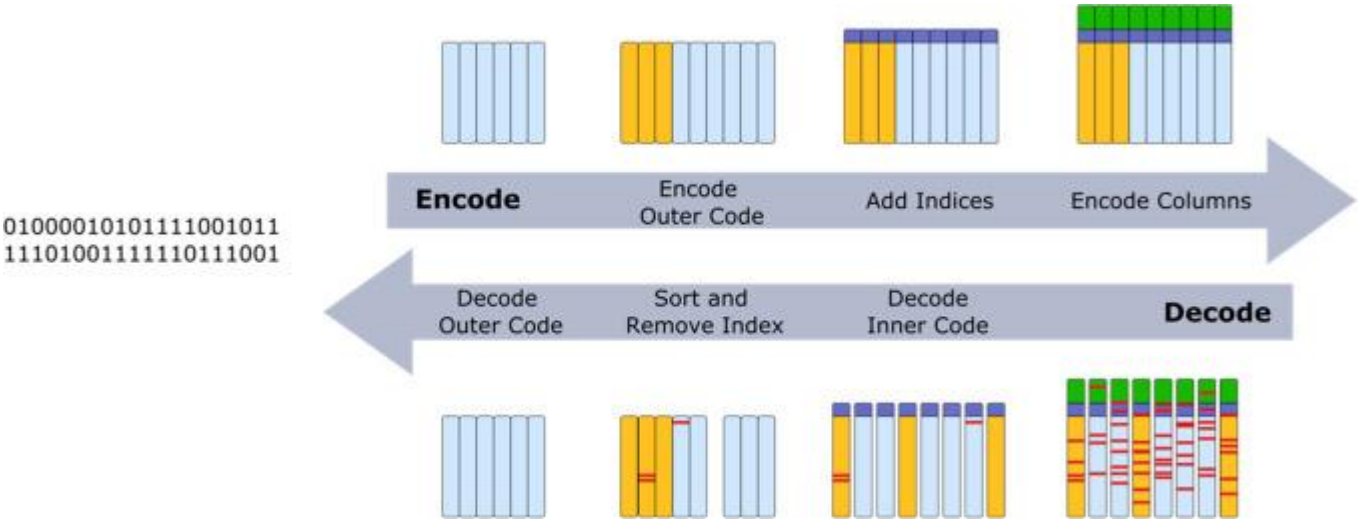
In addition to the errors during DNA sequencing discussed in the previous section, errors are also introduced during the synthesis, storage, and amplification steps of DNA data storage, which presents challenges regarding data decoding. While amplification and SBS-based platforms mainly introduce substitution errors (reading a C instead of a G, for example), synthesis dominantly causes deletions (e.g., missing a base) at a final rate of around 0.2–1% per nt.[20,62,70] Insertions (addition of extra bases) are uncommon and usually occur at less than 0.1% per nt, mainly because of synthesis.[20,62] In addition to biases in amplification efficiency, storage mainly contributes to shifts in the copy number distribution of the sequences, which leads to the unrecoverable loss of individual sequences over time, e.g., 8% after 94% of the DNA has decayed (i.e., four half-lives).[20,25] This means that, in general, sequence information is never recovered error-free. As the decoding of the stored data directly depends on this sequence information, both the loss of individual sequences and the introduction of errors into these sequences pose a risk on error-free decoding. While an increase of physical redundancy to cluster sequence information alleviates this problem, doing so is undesirable and inefficient because it drastically lowers the information density.[20] As considerations for cost and automation limit most of the potential for reducing error rates within the data storage workflow, sufficient redundancy must instead be implemented at the sequence level. Therefore, the presence of errors in DNA storage necessitates the use of principled coding/decoding algorithms. The goal of a good encoder/decoder pair is to enable perfect reconstruction from noisy data by introducing a minimal amount of logical redundancy. Error-correcting schemes tailored to DNA data storage consider that the written sequences are relatively short and typically stored in a spatially disordered manner. The optimal coding schemes depend on the noise profile of the storage system. Reliance on logical redundancy introduced by a combination of modern error-correction codes is sufficient for low error rates. However, both for low and large error rates, dominated by deletion errors, one also uses physical redundancy to recover the original information.[42]

An error-correcting code maps an original message to a larger one, which introduces redundancy. If this message is then sent over a noisy channel, thereby introducing random errors, these errors can be detected or corrected. A simple example of an error-correction code was used by Goldman et al.,[9] where each part of the information was written on four subsequent DNA sequences. Thus, the loss of sequences could be corrected if fewer than four subsequent sequences were lost. This coding scheme, however, was ill-suited for the used DNA channel because it had a low effective information rate, i.e., number of information bits per total number of encoded bits, and did not recover the whole message. In contrast, good error-correcting codes ensure data recovery with minimal redundancy. The maximal information rate that an error-correcting code can achieve is theoretically bounded.[89] This bound is known as the channel capacity and depends on the characteristics of the noisy channel. This means that the parameters of a good error-correcting code depend on the rates and type of errors. For example, the Reed–Solomon code can correct up to e erasures and s substitutions with $2s + e$ additional symbols.[90]

In 2015, a DNA data storage that used an error-correcting scheme, which enabled the recovery of full data, was realized by Grass et al.[4] Its encoding/decoding algorithm is explained in Figure 6. It uses an outer code that can correct for the loss of sequences, adds an index for each sequence to be able to retrieve the order of the sequences that are lost during storage, and uses an inner error-correcting code that can correct nucleotide errors within sequences. Following the original introduction of the inner-outer encoding scheme, the vast majority of subsequent works used such a scheme for DNA data storage.[14,42,44,70] In general, the outer code applies on the level of the original information, whereas the inner code protects single sequences or indices. However, different codes were used for the outer and inner codes. A Reed–Solomon code,[4,44,70] Fountain codes,[14] and LDPC (low-density parity check) code were used as an outer code.[91] As an inner code, a Reed–Solomon code was used by Grass et al. and Organick et al.[4,70] Blawat et al. proposed to protect the index separately with a bit-correcting code (BCH) as the inner code.[44] The inner–outer coding scheme works well for moderate error rates of 1–2% and substitutions. However, it cannot correct large error rates that are dominated by insertions and deletions. This is because no inner codes exist that work sufficiently well on short sequences in these noisy setups.[92] Here, the original message can be recovered by additionally exploiting the physical redundancy. For example, in Antkowiak et al.,[42] the noisy sequences were first clustered by similarity, then the information on multiple erroneous copies was combined to construct a sequence with fewer errors. This was achieved by an alignment step within the clusters and subsequent majority voting. This resulting sequence could then be sent through the usual decoding steps. Recent works have explored the development of efficient clustering methods tailored to DNA data storage, as well as efficient encoding schemes that allow the recovery of a sequence from multiple noisy reads.[93−96] Such codes could then be used as an inner code. This has led to a better understanding of efficient use of physical redundancy in DNA data storage. However, at this moment an optimal encoding/decoding scheme for long-term DNA storage, or even components of it, remains unknown. For example, the capacity of deletion and insertion channels or reconstruction from multiple reads with the combination of different errors

are not fully understood yet. Also, coding for these short unordered sequences remains challenging for very high error rates. Furthermore, different synthesis and sequencing techniques might motivate different approaches. For these reasons, error correction for DNA remains an active topic of research.

Figure 6.



Inner–Outer Code. ***Encoding***. The original information is first encoded with an outer code that introduces redundancy and protects against the loss of sequences. In Grass et al.[4] the original information was first grouped into blocks of multiple sequences (light blue). Then, each row was encoded with a Reed–Solomon code that adds redundancy (yellow). The columns correspond to single DNA sequences. These are labeled with a unique index (purple). Each column is then encoded with an inner code that adds logical redundancy on the level of each sequence (green). In general, the inner and outer codes need not add the redundancy separate from the original data, but instead return a modified longer word. ***Decoding***. The original information from the set of noisy sequences (errors marked in red) is retrieved by first decoding the inner code. This removes most errors within the sequences. For large error rates dominated by insertions and deletions, this step may be preceded by a clustering and alignment step that generates sequences with fewer errors from multiple noisy copies. The sequences are ordered by their index. The ordered sequences are then decoded by the outer code. Here, lost sequences correspond to erasures and erroneous sequences to substitutions. These are corrected by the outer code.

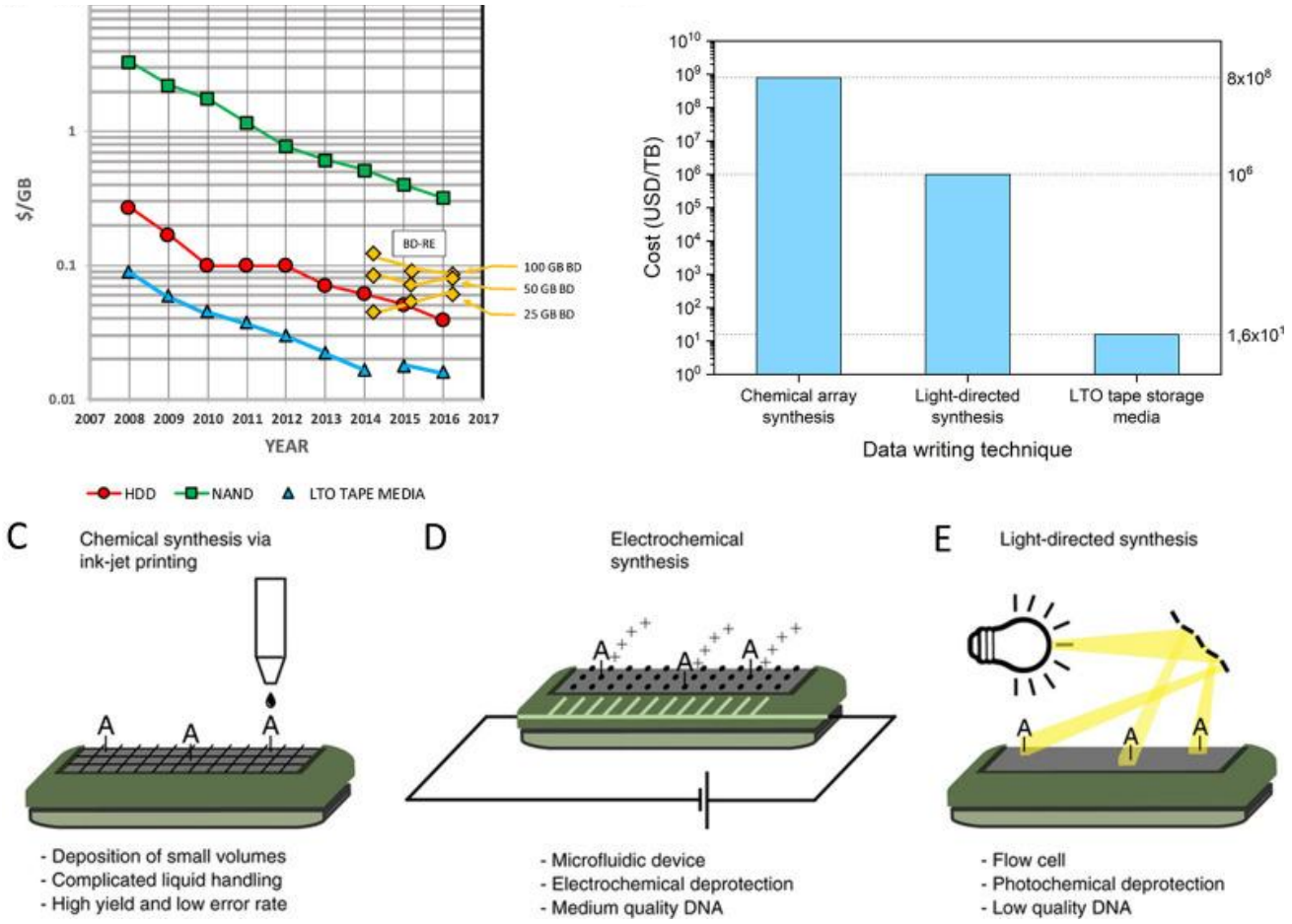
2.6. Limitations of DNA Data Storage

2.6.1. Issues Related to Cost

DNA has become a promising tool for next-generation data storage since it provides high data capacity and storage density[78] and it is possible to store it in multiple ways[37] over significant time periods.[4] However, in order to make DNA data storage standard, some limitations must be overcome. Arguably, the most important limit to the

development of DNA data storage is cost, especially in comparison with standard storage processes. Often, synthesis costs for DNA data storage are undisclosed;[12] however, it is possible to draw some conclusions about them. The synthesis of DNA oligos for data storage was column-based and was developed in the 1980s. Since then, this process has been fully automated, and now it allows the synthesis of 96–384 oligos simultaneously. The costs of this procedure range between 0.05 to 0.15 USD per nucleotide.[11] Array-based synthesis processes were developed in the 1990s. They lowered the costs because of their high-throughput nature, with an average price per nucleotide down to 10^{-4} USD. Thus, if a conservative estimate of 1 bit/nucleotide of encoded data is assumed, each terabyte of digital data would cost 800 million USD, on average.[12,97] In comparison, tape storage costs 7–8 orders of magnitude less, i.e., about 16 USD/TB of data, with prices decreasing by 10% every year (Figures 7A,B).[12,98] Considering this enormous disparity in cost between DNA data storage and magnetic tape, the outlook for DNA storage solutions initially appears dismal. That being said, DNA data storage has the potential to drop significantly in cost over time because of several key features. For example, optimized error-correcting codes could lower the cost[97,14] by increasing the overall efficiency of the storage process by means of accuracy reduction.[12] By capitalizing on error-correcting codes, it may be possible to work with cheaper, albeit less reliable, synthesis processes if it is assumed that any synthetic errors can be identified and corrected for upon readout, thereby leading to an overall reduction in cost. In 2020, Antkowiak et al. proposed that synthesis costs will drop to around 10^6 USD/TB (i.e., 2–3 orders of magnitude reduction) as a result of improved synthesis strategies, including large parallelization, optimization of reagents, and combination of nonvolatile DNA-based memories with logical operations (Figures 7B–E).[42] In addition, Antkowiak et al. estimated the marginal costs of the chemical synthesis of DNA. With the use of photolithography to synthesize 10 000 copies of each oligo, with a nucleotide reagent cost of 100 USD/g and a logical density equal to 1 bit/nucleotide, the cost of 1 TB of data stored in DNA would be $\sim 10^{-2}$ USD, with a chemical yield of 100%. Even if this chemical yield is impossible to achieve in industrial conditions, DNA data storage will be competitive against tape storage (20 USD/TB cost) even at 0.1% chemical yield. In the latter case, the cost of photolithographic DNA storage would be ~10 USD/TB, and synthesis conditions would be similar to the one used in surface chemistry (1000× reagent excess), which demonstrates that an optimization of chemical DNA synthesis processes is compatible with DNA data storage applications. Thus, Antkowiak et al. proved that the combination of synthesis processes that produce lower quality DNA oligos (i.e., photolithographic synthesis) and appropriate error-correction codes allows a major cost decrease in DNA data archives.[42] Regarding costs, there is also an important advantage with respect to traditional storage technologies that is worth mentioning. In fact, DNA storage systems' maintenance costs are expected to be lower than the ones of silicon devices in contemporary data centers.[97]

Figure 7.



(A) Cost trend of hard disk drives (HDD), NAND flash-based storage devices, linear tape-open tape cartridges (LTO tape), and optical Blu-ray (BD-RE). Image has been reproduced with permission under a Creative Commons Attribution 4.0 License (CC BY) from ref (99). Copyright 2018 AIP Publishing LLC. (B) Cost comparison between DNA synthesis for data storage and LTO tape storage. (C–E) Comparison of different DNA synthesis platforms and their characteristic traits. (C) Printing technology is primarily used by Twist and Agilent. (D) Electrochemical synthesis is employed by Custom Array. (E) Antkowiak et al. used light-directed synthesis. (C–E) Images reproduced with permission under a Creative Commons Attribution 4.0 License (CC BY) from ref (42). Copyright 2020 Springer Nature.

A strategy toward decreasing the costs of stored DNA data may be the enzymatic synthesis of DNA strands.[72] This synthesis could, in principle, decrease the costs of reagents even if the required enzymes are still rather expensive. It occurs in aqueous environments and it yields longer strands; however, error rates need to be assessed. A brief review of the principal trends in enzymatic synthesis is provided in section 2.1. The costs of enzymatic synthesis have been estimated by Jensen et al. for a template-independent enzymatic oligonucleotides synthesis (TiEOS) method.[100] The total costs of synthesizing 1000 strands of 1000 nucleotide length would be 136 USD with recycled TdT, 2700 USD by phosphoramidite technique, and 136 000 USD if a fresh stock of TdT

was introduced at every cycle. Thus, the costs of the enzymatic synthesis would be 1 order of magnitude lower than the phosphoramidite technique if the TdT was recycled.[100] The combination of advanced error-correcting codes and synchronization algorithms could possibly achieve lower costs of enzymatic DNA synthesis, as recently reported by Tang et al. This strategy allowed the enhancement of the coding rate to more than $\log_2 3$ per unit time and avoidance of deletions.[45] In the future, automation[39] of the reading, writing, and operative procedures, as well as the future developments of microfluidics, may forward DNA data storage toward a reduction of its economic costs.[12]

2.6.2. Issues Related to the Process Time Scales

Besides economic costs, automation could possibly lead to a reduction of the time costs for DNA data storage, as well. Indeed, the time requirements for the process are another limiting factor in the development of DNA data storage. For example, the reading speed is much lower than standard silicon-based storage media.[97] This could be detrimental, especially when the only possible alternative to retrieve a file would be to read the entire database: it would be a very slow process. For these reasons, DNA data storage systems have been proposed for long-term archival purposes[97] that need infrequent reading, while future investigations will be needed to fully realize random access.[78,75,70]

Conversely, in regards to nanopore reads of labeled DNA, each label is read in $[10^{-1}; 10^{1}]$ ms.[101,21,102]

The writing speed of DNA data storage is lower than that of standard technologies, too. The current writing speed for DNA archives is in the order of kilobytes/second, thus a reading/writing cycle has a significant cost in terms of time.[8] It is estimated that DNA data storage will need writing speeds in the order of gigabytes/second to be comparable with commercial cloud storage systems in around 10 years. This means DNA data storage must fulfill a gap of 6 orders of magnitude in regard to the writing (i.e., synthesis) and a gap of 2–3 orders of magnitude in regard to the reading (i.e., sequencing).[12]

In order to enhance the read/write speed of DNA data storage, one of the goals should be to make it suitable for frequent data reads and modifications. This is another pivotal reason for the investigations about synthetic polymers as data storage tools, together with the mentioned high cost of DNA.[97]

While writing and reading operations regarding DNA-stored data need to be improved, when it comes to preservation time, DNA is better than current storage technologies. Indeed, the maximum preservation time of information is 50 years for digital memories and 500 years for paper, while it is millennia for inorganic matrix-encapsulated DNA.

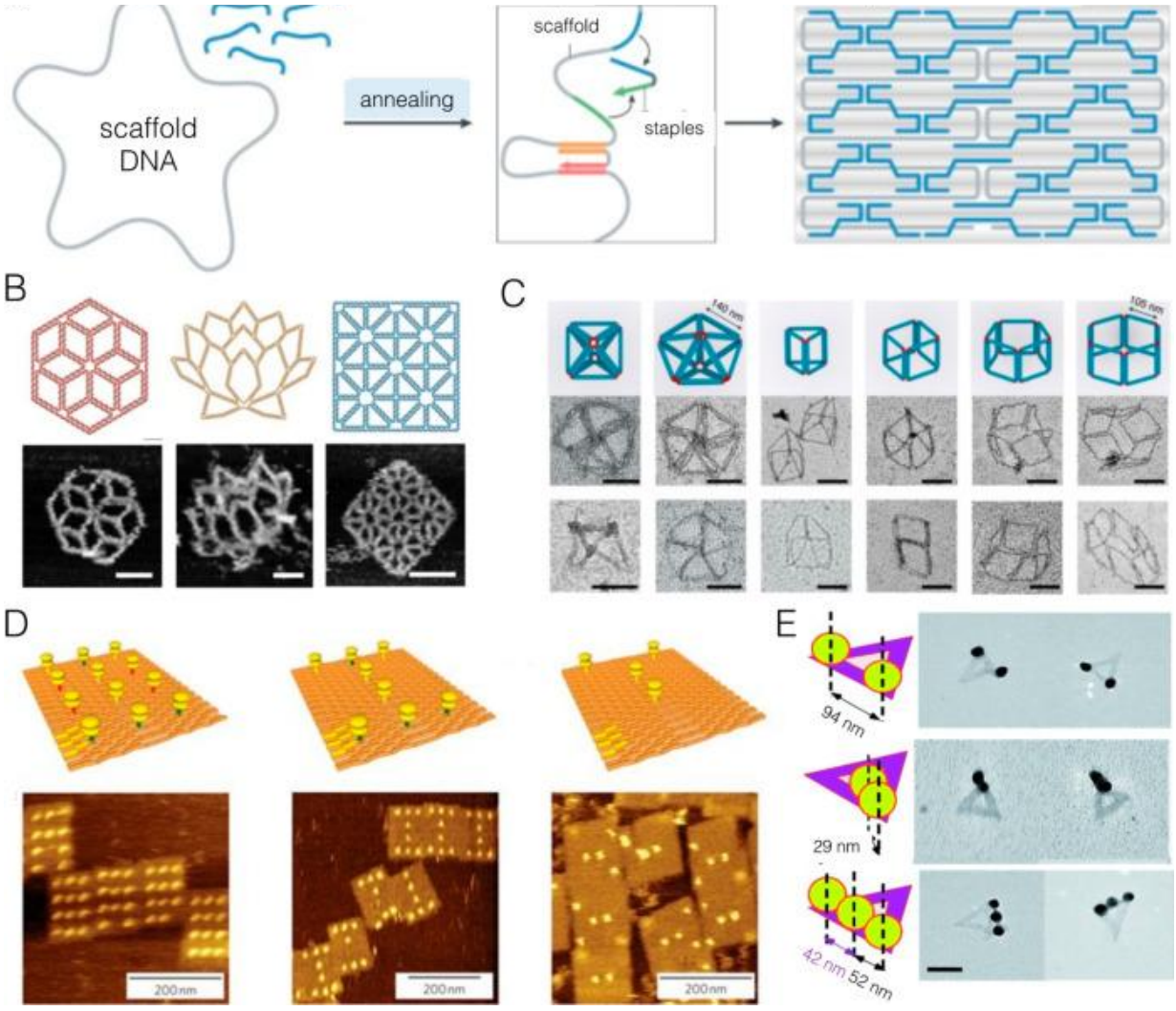
In conclusion, DNA data storage presents both advantages and disadvantages with respect to traditional storage methods regarding costs. It is also for this reason that research interest is growing in this field.

3. Structure-Based DNA Data Storage

3.1. DNA Nanotechnology Versus Synthetic DNA Sequence for Digital Data Storage

DNA nanotechnology may also be employed to overcome the limitations illustrated above in synthesis and reading. Because of the self-assembled nature of DNA nanostructures (Figure 8), it is possible to significantly reduce the synthetic demand and to eliminate the need for next-generation sequencing for DNA data storage. DNA nanotechnology leverages the unparalleled molecular recognition motifs of the nitrogenous bases to create arbitrary two- and three-dimensional structures from the self-assembly of user-defined DNA strands.[24,103] Through careful design of the sequences of these strands, which can be easily synthesized in an automated manner or even purchased from commercial vendors, exquisite control over their final assembly can be realized, thereby enabling the construction of nanoscale shapes and patterns. The main approaches in structural DNA nanotechnology can be divided into three groups: DNA origami, DNA tile assembly, and wireframe DNA structures,[104] all of which have been extensively reviewed elsewhere.[103] Among these, DNA origami is the most widely used method for the construction of DNA-based data storage structures at the nanoscale. Importantly, all of these bottom-up approaches enable the production of asymmetric patterns, which is a key criterion for data storage applications: instead of encoding information directly into the sequence of bases, data may be stored in the three-dimensional shape of these assemblies.

Figure 8.



DNA nanostructures are data storage architectures. (A) DNA origami leverages the specific base-pairing motifs of DNA to create arbitrary structures. When a long scaffold strand (several thousand nucleotides in length) is combined with hundreds of short "staple" strands, complementary regions on the different strands will hybridize, thereby folding the scaffold into a desired conformation. These structures can then be examined using (B) atomic force microscopy or (C) electron microscopy, for example. (D) Data can be written onto DNA origami sheets through the site-specific addition of proteins; the data may be read using AFM. (E) Nanoparticles can also be controllably positioned on DNA origami with nanometer-scale resolution, which enables data writing with cryo-EM readout. (A) Image reproduced with permission from ref (108). Copyright Springer Nature 2021. (B) Image reproduced with permission under a Creative Commons Attribution 4.0 License (CC BY) from ref (109). Copyright 2019 AAAS. (C) Image reproduced with permission from ref (110). Copyright 2020 Springer Nature. (D) Image reproduced with permission from ref (111). Copyright 2010 Springer Nature. (E) Image reproduced with permission from ref (112). Copyright 2010 Wiley-VCH.

Because of the noncovalent nature of DNA nanostructures, they can be reconfigured using established strategies, including strand displacement,[26] thermal annealing,[105] and pH changes.[106] The reversible Watson–Crick base pairing means that, unlike data encoded directly into the primary DNA sequence, data storage platforms based on DNA nanostructures can be "erased" and "rewritten" multiple times without requiring any laborious chemical synthesis, which decreases the synthetic demand and cost associated with these methods.[25] Additionally, the reconfigurable nature of these constructs enables their use in data operations and computation, analogous to existing computer memory systems. Because each bit is formed through self-assembly, it is also possible to encrypt information by initially omitting a key element from the assembly mixture; only upon addition of the correct "password" molecule can the DNA-based data be "read."

Compared with encoding data within the nucleotide sequence itself, data storage based on DNA nanotechnology has one major drawback: data storage density. While data written directly into the DNA sequence theoretically allows 1 exabyte (or 1 billion gigabytes) to be stored in every cubic millimeter of DNA,[107] the data density that has been attained so far using DNA secondary structure is much lower because it requires ~100 base pairs per bit.[25] That being said, this density is still approximately 3 orders of magnitude higher than current hard drive technologies, with further improvements conceivable through the optimization of the 3D DNA structure. Considering the advantages of encoding information into the secondary structure—including ease of readout, synthetic simplicity, and reconfigurability—this is a minor obstacle and one that may be mitigated through the careful design of DNA nanostructures.

3.2. DNA Nanostructure-Based Information Storage Platforms: Assembly and Readout

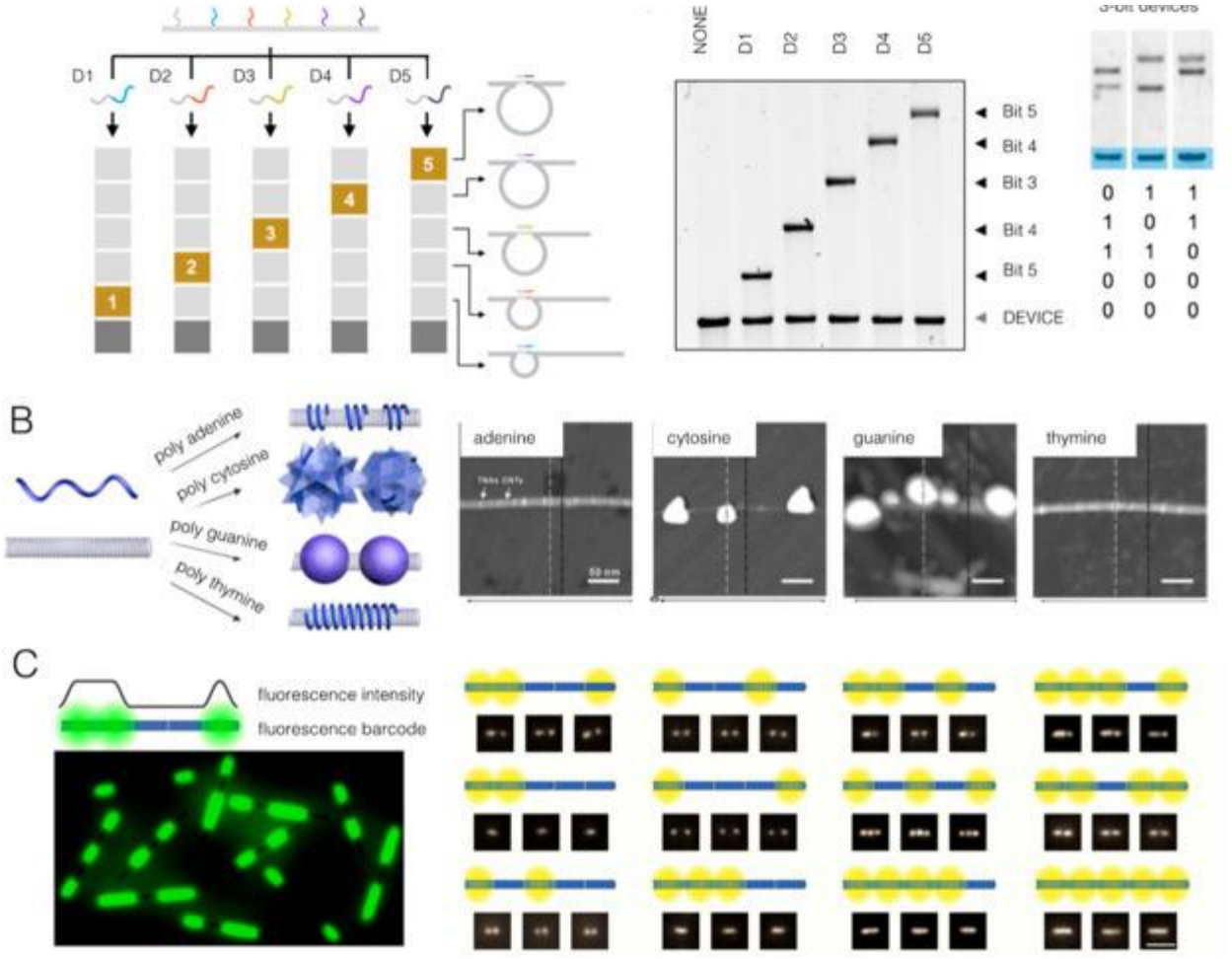
When comparing DNA nanostructure data storage to traditional sequence-based methods, the major differences lie in the reading and writing steps. In particular, standard DNA data storage requires slow and costly DNA synthesis, while DNA nanostructures already store molecular data in two- and three-dimensional objects. In fact, the assembly of DNA origami is, itself, a molecular information encoding process, wherein the long scaffold strand is folded with hundreds of short "staples" to form a predetermined structure (Figure 8). The size and morphology of the resulting structures can be assessed using various ensemble and single-molecule characterization methods, thereby enabling the readout of information stored in the shape and structure of these nanoscale assemblies. The use of this suite of techniques (described in detail in the following sections) has two major advantages: (1) Depending on the design and the physical attributes of the data storage structure, it may be possible to perform more than one type of characterization. Comparing the results of different readout methodologies may allow for the identification of systematic biases in each modality, which generates a feedback cycle wherein structures may be improved upon and recharacterized. (2) The identification of larger structures (on the order of approximately tens of nanometers) de facto requires lower resolution than the differentiation of single

bases, thereby facilitating the use of less precise techniques without sacrificing accuracy. Additionally, because the single-molecule readout methods used for the assessment of DNA nanostructures are also used in DNA sequencing, these techniques are constantly improving: in this way, the advancement of sequence-based DNA data storage also supports the growth of alternative, structure-based approaches.

3.2.1. Gel Electrophoresis

A first and very simple method to read data is the use of gel electrophoresis, which remains one of the key methods to differentiate DNA nanostructures of different shapes and sizes, as well as to assess their yield. Through the formation of DNA nanostructures with prescribed differences in size, it is possible to encode information and then read this out using the discrete bands formed on a gel. To this end, simple structures involving hairpins, loops, or G-quadruplexes placed along linear DNA backbones can also be used to store digital data. For example, Halvorsen and Wong used the change between a closed loop structure (“1”) to a linear structure (“0”)—which have different elution times by gel electrophoresis—as a binary switch. The authors used electrophoresis to demonstrate the readout of an 11 byte ASCII message.[113] The creation of many loops of different sizes, each distinguishable by gel electrophoresis, offers a greater number of possible bits in each lane (Figure 9A).[114] The formation of loop structures is not the only operation of DNA nanostructures that can be directly probed using gel electrophoresis. In an alternative approach, five single-stranded nucleotides were annealed together to form an assembly with three addressable overhangs; when complementary strands to each of these overhangs were introduced, the site changed from a “0” (single-stranded) to a “1” (double-stranded) state, which could then be reversed using strand displacement.[115] These examples highlight the simple and inexpensive nature of gel electrophoresis as a readout platform, especially when compared with optical, electrochemical, and AFM-based methods. However, the relatively long read times and low data capacity of these methodologies limit their applicability. Gel electrophoresis, being a bulk measurement, also requires substantial quantities of DNA for readout relative to single-molecule methods like AFM, electron microscopies, and nanopore techniques.

Figure 9.



Examples of DNA nanostructures for digital information storage. (A) The folding of DNA origami into loop structures upon binding of a biomolecule target generates a shift in the assembly's electrophoretic mobility. Image adapted with permission under a Creative Commons Attribution 4.0 license (CC BY) from ref (114). Copyright 2017 Oxford University Press. (B) The association of different DNA sequences to carbon nanotubes produces an array of morphologies and, therefore, can be used to produce barcodes. Image adapted from ref (116). Copyright 2019 American Chemical Society. (C). Data strings based on regions of varying fluorescence intensities along a DNA nanotube can be read out using single-molecule fluorescence microscopy. Image adapted from ref (117). Copyright 2021 American Chemical Society.

3.2.2. Fluorescence

Bulk fluorescence measurements can read out data encoded into DNA nanostructures. In an early example, DNA strands were used as "molecular memory" by transitioning thermally between a hairpin structure (unwritten state) and a duplex structure (written state).[118] The oligonucleotides were appended with fluorophore/quencher pairs; as the thermal cycling occurs, the fluorescence output reversibly switches between two defined states to produce a binary signal. Unfortunately, because this process is performed in

solution, the whole memory is erased simultaneously, which highlights the need for alternative strategies that enable spatial addressability. To this end, single-molecule fluorescence methods may be used instead to read out DNA origami breadboards appended with fluorophores. In one approach, termed “polychromic address multiplexing,” DNA origami was separated into spatially resolved “cells,” each of which contained a set of fluorophores appended to DNA. Some of these linkers contain photocleavable groups, which enables the disruption of energy transfer processes between adjacent dyes, thus resulting in a fluorescence change. The switch between two possible intensity values provides the binary logic in this system.[119] Through the use of single-molecule total internal reflection fluorescence (TIRF) microscopy, it is possible to decode fluorescent barcodes assembled on DNA nanostructures.[120] Pan et al. utilized this diffraction-limited imaging technique to devise a method to group fluorophores into bright (“on”) lengths along a DNA origami rod.[121] Such bright spots were separated by dark (“off”) regions to create geometric barcodes using only one color of emitter (Figure 9C). Another tactic used a DNA origami “breadboard,” which was divided into a grid of pixels or an “indexed matrix of digital information.” Each specific location on the origami represents a bit, with the presence (“1”) or absence (“0”) of a docking site for a fluorophore encoding binary information.[22] Docking sites are located using DNA points accumulation for imaging in nanoscale topography (DNA-PAINT), a form of super resolution fluorescence imaging that relies on transient binding of short DNA strands to prepositioned sites on an origami structure.[122] In this example, unique data patterns are created by selecting which staple strands within the origami possess data domains. This approach also uses error-correction algorithms that enable message recovery even when individual docking sites are missing. Unlike DNA sequencing, which requires multiple reads to reach a consensus, this tactic can read 750 origami to reach a 100% probability of full data retrieval, which means that only femtomoles of material are needed.

3.2.3. Atomic Force Microscopy

Early examples of DNA origami were reported in the mid 2000s and involved the assembly of 2D arrays to form various images, including the letters of the alphabet,[123] a nanoscale Mona Lisa,[124] and a map of the Americas.[125] Atomic force microscopy (AFM) was used to “read-out” images formed by DNA origami, and this remains a key technique for the study of DNA-based nanomaterials.[126] AFM measurements detect differences in height over a sample surface, without affecting the sample, thus rendering this method ideally suited to reading out three-dimensional patterns on DNA origami. Binary information can be written by precisely placing nanoparticles or proteins at defined positions on a DNA breadboard. In the context of DNA data storage, Zhang et al. demonstrated in 2019[127] a “DNA braille” system, which was prepared by patterning biotinylated overhangs onto DNA origami. The data in this system are encrypted; only when streptavidin is added and binds to biotin does the pattern become readable by AFM. The decryption time for this method is 1–2 h, including sample processing, imaging, and readout—this time could be reduced by using high-speed AFM methods and fully automated image analysis algorithms. Similarly, Fan et al. used AFM to decode information stored in DNA domino arrays.[127] The use of DNA overhangs bearing

streptavidin enables the use of strand displacement reactions to controllably erase and rewrite data on the DNA origami surface,[128] thereby underlining the advantages of DNA nanotechnology as an information storage platform. AFM is also suitable to look at DNA positioned on other types of nanomaterials; for example, it was found that condensing DNA strands onto carbon nanotubes creates height differences that were observable by AFM (Figure 9B). Control of the patterning of these protrusions, which interestingly do not rely on DNA hybridization, may allow for the production of two-dimensional barcodes on carbon nanotubes.[116]

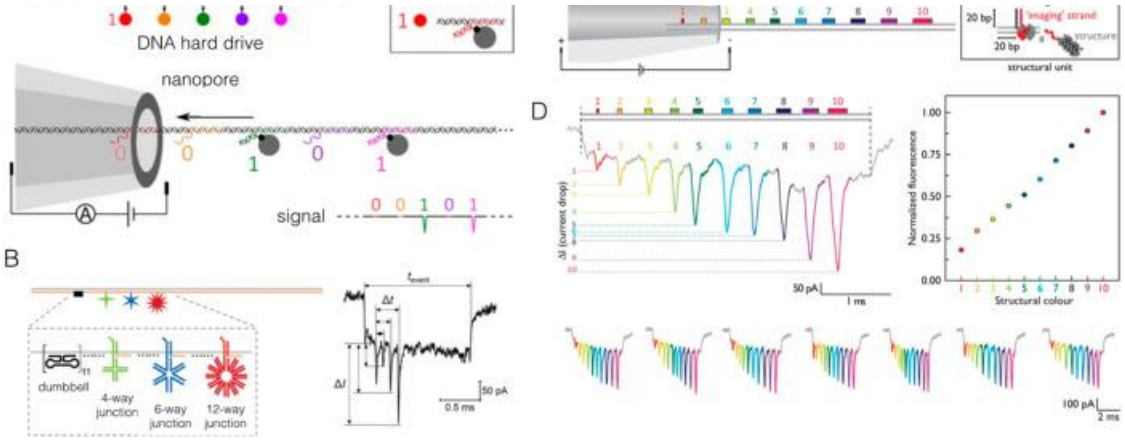
3.2.4. Electron Microscopy

Relying on similar principles, the decoding of DNA nanostructures can also be achieved using electron microscopy (EM). DNA itself can be difficult to visualize using EM because of insufficient electron density-related contrast, and therefore, often requires staining. As such, EM is better suited to the examination of hybrid structures, wherein the DNA is used to create “barcodes” made of gold nanoparticles,[129] for example. Different barcodes can then be used to track the cellular uptake of various nanostructures because EM allows for the identification of subcellular compartments. EM exhibits some of the same advantages and pitfalls of AFM: while these techniques allow for high-resolution two- and three-dimensional images to be formed of DNA nanostructures, they are time-consuming and expensive, as well as relatively low-throughput. As cryo-EM and liquid-cell EM techniques continue to improve, the direct imaging of biomolecules might offer an alternative in the future with better resolution on the single-molecule level even without the use of staining or nanoparticles.

3.2.5. Nanopore Measurements

More recently, through the use of long DNA backbones as in DNA origami, the organization of DNA protrusions has been used to produce three-dimensional DNA barcodes[21] or hard drives that may be read using solid-state nanopores. Nanopore methods require no labeling for readout, which makes them an attractive alternative to fluorescence. Briefly, an electric field is applied across a nanoscale hole (made from glass or Si_3N_4, for example), which causes molecules to translocate through this nanopore. As the analyte passes, it modulates the ionic current signal because of its 3D shape blocking the pore—in this way, the structure of the DNA nanoconstruct is translated directly into an electrical signal (Figure 10). The resulting current–time traces can then be analyzed using automated methods, which allows for rapid data decoding.

Figure 10.



DNA data storage structures relying on nanopore readout. (A) An encrypted “DNA hard drive,” wherein readout may only occur once the correct molecular “keywords” have been added. Streptavidin molecules (gray circle in inset) partially block the nanopore as they translocate, which causes a momentary decrease in the current. Image reproduced from ref (25). Copyright 2020 American Chemical Society. (B) Multilevel barcoding is achievable by exploiting DNA junctions with different sizes, which create current drops of variable magnitude. Image reproduced with permission under a Creative Commons Attribution 4.0 License (CC BY) from ref (102). Copyright 2021 Wiley-VCH. (C) A DNA barcode with “structural colors” can also be formed by closely packing structural units, which therefore read as one protrusion. These units may be based on either monovalent streptavidin or a DNA cuboid. (D) Nanopore microscope can be used to detect up to 10 structural colors within the same DNA data string. The correct identification of the “color” was verified using fluorescence microscopy, wherein fluorescently labeled (5′-fluorescein) structural units were used. (C,D) Images reproduced with permission under a Creative Commons Attribution 4.0 License (CC BY) from ref (130). Copyright 2022 Springer Nature.

The use of nanopores to read out digital information encoded in DNA nanostructures was demonstrated by Bell and Keyser, who fabricated “DNA barcodes” to capture proteins.[21] The authors used conical quartz nanopores with diameters of ∼14 nm for a 3-bit barcode that could be assigned with 94% accuracy. Now, these quartz nanopores can read out DNA hairpins along a carrier strand with a density of approximately 1 bit per 30 nm—ca. 3 times the data density of conventional hard drives.[25] One of the major benefits of this method is their high speed: a single “DNA hard drive” can be read out on the millisecond time scale using a quartz nanopore because of the superior signal-to-noise ratio when compared with DNA sequencing. Solid-state nanopores combined with DNA nanotechnology have since been used to save and encrypt a grayscale image.[102] Streptavidin-labeled scaffolds can also be used to create a secure data storage system that requires the correct molecular “keywords” to decode the data within the structure (Figure 10A). Multilevel storage architectures have been achieved using different DNA junction sizes to create a quaternary encoding system (Figure 10B).[102] Increased storage density beyond binary barcodes can also be achieved by creating blocks of repeating structural units that appear as a single

protrusion within the nanopore, which creates “structural colors” to generate up to 10 data levels.[130] Compared with fluorescence, sequencing, or gel electrophoresis-based strategies, single-molecule nanopore measurements require less material and enable faster data reading; through a combination of this technology with deep learning methods,[131] real-time nanopore data analysis is attainable.

Another important feature is random access, as demonstrated in 2021 by Bošković et al.[101] In their work, random access of DNA barcodes was performed by exploiting a modified PCR method to increase the number of the target DNA nanostructures. Indeed, DNA structural barcodes were annealed as short oligonucleotides containing protrusions on single-stranded DNA (ssDNA) scaffolds to form digital bits at precise locations. In these structures, DNA nicks were ligated to favor the copy of the barcode by PCR. Each of these structures had a noncomplementary end, which acted as a barcode-specific primer template for the random access of data.

3.2.6. Alternative Approaches and Polymer Chemistries

The use of double-stranded DNA as a storage medium was also exploited in recent work by Tabatabaei et al.[72] on DNA punch cards. This macromolecular storage technology was used to encode the information in the sequence of bases of the DNA strands by using their sugar–phosphate backbone, i.e., topologically. Indeed, a pattern of nicking positions was precisely realized on the backbone of native dsDNA, and here, information was encoded by means of absence (i.e., 0) or presence (i.e., 1) of nicks. On the basis of enzymatic modification of DNA, nicks enable adding several functionalities to the storage system, for example, single-bit random access, pooling, and in-memory computation. However, the DNA punch cards system was able to store only up to 14 kB of digital information. Therefore, additional research is foreseen toward scaling its costs.

DNA as a natural polymer is not the only solution for data storage technologies. Therefore, researchers started to look for alternative molecular storage platforms based on synthetic polymers. Synthetic polymers can be used to increase stability against chemical degradation while offering a wide range of base modifications. Although alternative DNA bases have been introduced, synthetic polymers could be prepared using a set of monomers with a wider set of codes which expands the alphabet for data encoding.[97] First experiments reading single-stranded synthetic biopolymers indicate that the reading step can be performed with biological nanopores without the use of an enzyme slowing down the translocation.[132] As an example, Cao et al. used informational biopolymers composed of a backbone of poly(phosphodiesters) with dideoxyadenosine at both ends, and engineered-aerolysin nanopores. The results suggest a path to single-bit resolution at least in short polymers, however machine learning and training are needed for the successful readout. The study suggests an alternative way to store information with high density. The idea to use the backbone of an organic polymer to store digital information is similar to the approach discussed for DNA nanostructures.

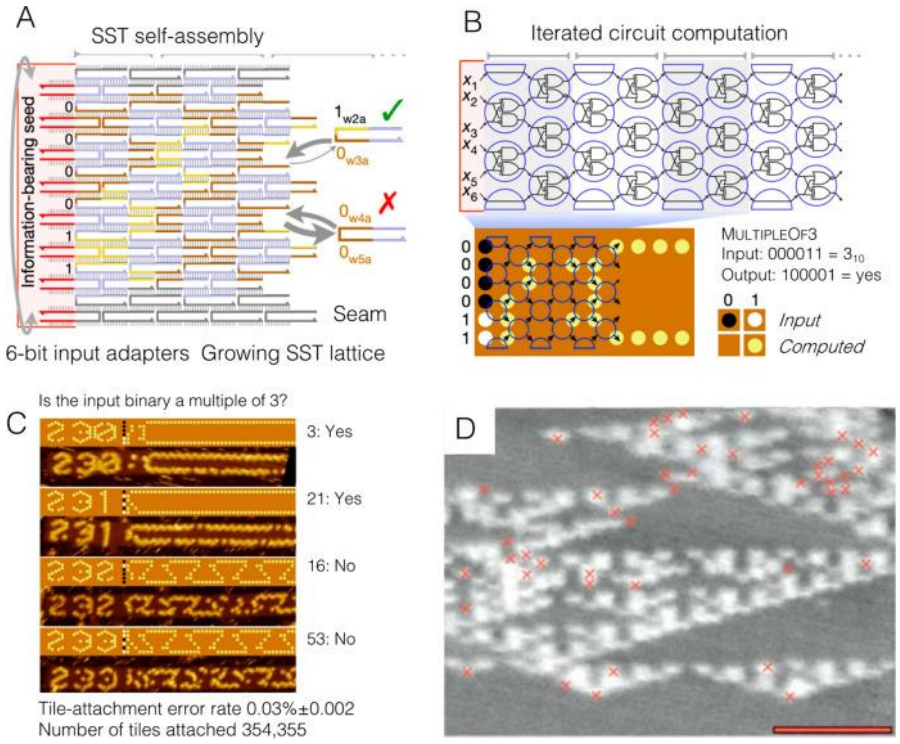
Apart from DNA, other organic molecules have been recently proposed.[133,134] Two interesting examples are the use of peptide sequences for data storage, as reported by Ng et al.[133] and the use of urethanes as reported by Dahlhauser et al.[134] Unfortunately in both these cases, reading required the use of mass spectroscopy, with the consequent limitation in terms of costs and speed. Recent advances in nanopore-based readout of short peptide sequences[135] may speed up developments in this area.[53]

3.3. DNA Nanotechnology for Molecular Computation

The storage of data in DNA is undoubtedly an exciting possible solution to our ever-expanding data storage needs. This technology may lead to future hybrid electronic–biomolecular computing systems in which some portion of the burden of data storage is supported by DNA encoding, which raises the question: “Can more of the computer system’s functions be carried out using DNA?” By reducing the time overhead of conversion to a digital format and directly undertaking data processing tasks with DNA-based computation, it may be possible to create molecular computing systems that are more efficient than conventional electronic analogues. Because of the noncovalent nature of DNA nanostructures, these materials are primed for use in molecular computation. A working prototype for a DNA computer was developed by Adleman in 1994,[136] wherein he used a separation-based approach to calculate a Hamiltonian path in a graph with seven summits. This problem was particularly suited to a molecular computing approach because it is an NP-complete problem; while verification of a putative solution has a complexity that is linear with respect to the number of nodes, the path space search is exponential in complexity with respect to the same. In a DNA computer, however, each DNA molecule plays the part of a separate processor, which enables many parallel operations to be carried out in a small reaction volume. This strategy greatly accelerates the initial path search, as statistics predict that DNA constructs corresponding to every possible path should be produced upon mixing. The task is then reduced to one of selection and filtering by removing invalid paths. The Adleman experiment acted as a proof of concept: in practice, the process was more time- and labor-intensive than a conventional digital approach. Nonetheless, the possibility of a DNA-based computer inspired researchers to further develop Adleman’s method and to devise advanced and powerful general DNA computing solutions. Early experimental and theoretical work examining the possibilities of DNA computation was focused on this parallelization and the benefits that it offered with regard to efficiently solving other NP-complete problems.[137,138] Recent work on DNA computation has moved away from such problems toward recreating deterministic logical operations, for example, addition[139] and multiplication,[140] with definite outcomes. Su et al. produced DNA logic cascades, which allows the buildup of a full adder, a 4:1 multiplexer, and then, they combined these with other logic circuits to produce a DNA arithmetic logic unit (ALU): the foundation of general-purpose processors.[139] These applications demonstrate the methods that can be used to mitigate error in DNA computation, which arise from the leeway and tolerance of mismatch inherent in sequence-specific DNA hybridization.

Larger engineered DNA nanostructures show great promise for use in biomolecular computation as well as small origami structures. Robust, rigid DNA tiles with programmable “sticky ends” have been made using double-crossover (DX),[141] triple-crossover (TX),[142] and single-stranded tile (SST)[143] motifs and used for a variety of algorithmic self-assembly experiments. This is facilitated by the logical equivalence of these tiles with Wang tiles, which are theoretical constructs with specified interactions that can simulate a Turing machine. A correctly designed, self-assembling set of these tiles is theoretically able to perform any computation that can be carried out by a conventional computer. Past applications of this idea include the design of a set of TX tiles that carry out a cumulative XOR operation, a set of DX tiles that self-assemble into a Sierpinski triangle, and impressively, a set of 355 SSTs that can be used to produce a variety of cellular automata capable of carrying out a number of computational tasks (Figure 11).[144] Particularly interesting in the latter example is the ability to controllably reintroduce indeterminism by including a plurality of tiles that could fill a given niche and leaving the ultimately realized pattern up to competition. This brings about a marriage of the benefits of deterministic logic and the power of indeterministic computing to solve combinatorial problems, thereby highlighting the utility of DNA nanotechnology not only for data storage but also for molecular computation.

Figure 11.



Tile-based computations and algorithmic self-assembly. (A) Self-assembly by SSTs. From a seed, tiles attach to the frontier of a growing SST lattice according to interaction rules determined by their exposed recognition sequences. (B) An iterated Boolean circuit mimicking the function of a computation to determine whether or not a binary number is a multiple of 310. A long enough lattice will settle into one or another fixed

pattern corresponding to the calculation result. (C) The result of four “multiple of 3” tilings. The numbers at the left mark the experiment number. The tilings correctly determine which input numbers have a factor of 3. (A–C) Images adapted with permission from ref (144). Copyright 2019 Springer Nature. (D) A Sierpinski triangle created by a cumulative XOR computation performed by DNA tiles. Sierpinski’s triangle is a fractal pattern, and the self-assembly rule that creates it is Turing complete. Images reproduced with permission under a Creative Commons Attribution 4.0 License (CC BY) from ref (145). Copyright 2004 PLoS Biology.

4. Conclusions and Outlook

DNA data storage—both the sequence- and structure-based versions—offers the possibility of storing digital information at very high data density. This promise has led a large number of actors (public and private institutions, corporations, etc.) to invest on the quest for advanced methods and experiments. Although great advances have been made toward DNA data storage, it is not yet competitive against conventional storage technologies. Significant challenges need to be overcome, in particular regarding writing speed and, hence, cost. While stored data size has been markedly increased, the current record for DNA digital data storage is still around 200 MB, with single synthesis runs lasting about 24 h.[8,12] Achieving the storage of TBs of data at a low cost is unattainable with the current techniques. Toward this goal, great efforts on the development of encoding schemes, writing and reading processes, and storage procedures are presently being made.[146,9,81,44,74,14]

As the chemical and enzymatic processes for making sequence-defined nucleic acids continue to improve, the cost and time associated with writing DNA-based information is continually decreasing. These improvements are particularly important for sequence-based storage, but they importantly reduce costs for structure-based approaches, as well. Additionally, as alternative chemistries emerge, including unnatural nucleotides[147] and small molecules that can modulate the structure of DNA,[148] the parameter space for structure- and sequence-based DNA data storage is continually expanding. Importantly, these chemistries not only widen the breadth of materials that can be produced, but also may further extend the lifetime of DNA sequences, as these modifications render DNA less recognizable to enzymes.

For data readout, DNA sequencing is rapidly advancing, but current methods would be incompatible with unnatural monomer units, which limits the scope of the methods. Furthermore, all current DNA sequencing techniques require molecular machines like polymerases, which set fundamental limits for the throughput per enzyme, thereby meaning there is an upper threshold on the rate of sequencing even with massive parallelization. Emerging, rapid approaches to establish polymer sequence or three-dimensional structure one molecule at the time will improve the competitiveness of DNA data storage. Both natural and chemically modified oligonucleotides, as well as hybrid nanostructures involving DNA and quantum dots or nanoparticles, may be read out using solid-state nanopores. The versatility of this methodology, which hinges on the

possibility of finely tuning nanopore size, makes this an attractive avenue for the future characterization of both pure DNA and composite materials. Through the use of quantum dots or fluorescent dyes, nanopore readout may be also combined with optical techniques to reduce the readout error rate without requiring enzymes to slow translocation.[149] While the use of higher order nanostructures or composite nanomaterials does sacrifice data density, the advantages of these methods are expected to outweigh this drawback. In terms of synthesis, DNA nanotechnology greatly simplifies assembly procedures and produces structures that can easily be reconfigured. Indeed, computation is a natural extension for DNA nanotechnology, especially considering the vast library of naturally evolved enzymes that nature uses to copy, change, and repair genetic information. The interface between these natural systems and DNA nanotechnology is an active area of research, which generates other possibilities for DNA data storage that leverage nature's evolved machinery. We foresee that DNA nanostructures made for information storage will find audiences in cryptography, steganography, and other fields[4,58,116] and that combining DNA data storage with data analysis techniques such as neural networks will afford opportunities in a growing number of sectors.[127]

Because of the long-term stability of DNA under appropriate storage conditions, we predict that archival storage will be the most valuable application for DNA data storage. In this cold storage setting, information would be infrequently accessed from a relatively static DNA database. Considering that long-term, archival storage[97] operates over long time scales—decades, centuries, and possibly millennia—this application requires only infrequent access to the stored information, which substantially reduces the impact of reading costs and long read times associated with DNA data storage. While the long-term stability of DNA, itself, is firmly established, further studies on the lifetimes of noncovalently assembled DNA nanostructures will need to be conducted to ensure that data stored in these formats are not compromised over time. Specifically, encapsulation and retrieval of DNA nanostructures in silica beads and other matrices should be examined, as well as the readability of DNA nanomaterials after prolonged freezing. It is also important to mention that the preservation of DNA digital archives can be implemented using not only in vitro substrates but also in vivo approaches.[150−153] As three-dimensional nucleic acid nanostructures have also successfully been produced inside cells,[154] there is potentially important synergy between in vitro/in vivo DNA nanotechnology and data storage, which remains, as of yet, unexplored.

Even in the context of archival storage, a DNA database, like its electronic analogues, would benefit greatly from dynamic properties that allow data to be erased, rewritten, and updated. For example, in-storage file operations and computations, as well as the ability to repeatedly access DNA databases, would reduce DNA synthesis costs and abrogate the need to store multiple copies of archives. In this area, DNA nanostructures may present advantages over traditional sequence-based storage methods, as the reconfiguration of these supramolecular moieties is firmly established, though rarely in the context of information storage. The implementation of dynamic properties and a full

characterization of the kinetics of these processes would bring DNA-based storage systems one step closer to practical viability.[78]

A combination of sequence- and structure-based approaches could represent a significant advancement to overcome the various hurdles associated with DNA data storage. For this field to reach its full potential, cooperation between scientists from a range of research areas will be essential to produce the advanced chemical techniques, instrumentation, characterization methods, and automated analysis tools that are required. As the wide range of topics, from mathematics to polymer chemistry, shows, data storage based on polymers will demand multidisciplinary consortia that ideally design the whole process from data encoding to decoding with a bottom-up approach.

Glossary

Vocabulary

sequence-based DNA data storage
: storage approach in which information is stored in the nucleotide sequence of many individual DNA strands

structure-based DNA data storage
: storage approach in which DNA is designed in a way that allows information to be stored in its structural features, e.g., 2D and 3D shape

data encoding with error-correcting codes
: conversion from digital, binary data into the primary sequence (sequence-based) or 2D/3D structure (structure-based) of DNA by adding redundancy to counteract errors and partial data loss

random access
: process by which a certain subset of information is selected (e.g., a single file) from a large pool of information

reading
: process by which the data encoded in DNA are read; this process varies according to if the DNA data storage is sequence-based or storage-based

decoding
: conversion of the information that was read from DNA into binary data; this conversion happens both in sequence-based and in structure-based DNA data storage

DNA storage: research landscape and future prospects

Yiming Dong [1,b], Fajia Sun [2,b], Zhi Ping [3], Qi Ouyang [4,5], Long Qian [6,✉]

PMCID: PMC8288837 PMID: 34692128

Abstract

The global demand for data storage is currently outpacing the world's storage capabilities. DNA, the carrier of natural genetic information, offers a stable, resource- and energy-efficient and sustainable data storage solution. In this review, we summarize the fundamental theory, research history, and technical challenges of DNA storage. From a quantitative perspective, we evaluate the prospect of DNA, and organic polymers in general, as a novel class of data storage medium.

Keywords: DNA storage, information, compilation, sequencing

INTRODUCTION: INFORMATION AND STORAGE

Human civilization went through paradigm shifts with new ways of storing and disseminating information. To survive in the complex and ever-changing environment, our ancestors created utensils out of wood, bone and stone, and used them as media for recording information. This was the beginning of human history [1]. With the development of computer technology, the information age has revolutionized the global scene. Digital information stored in magnetic (floppy disks), optical (CDs) and electronic media (USB sticks) and transmitted through the internet promoted the explosion of next-generation science, technology and arts.

With the total amount of worldwide data skyrocketing, traditional storage methods face daunting challenges [2]. International Data Corporation forecasts that the global data storage demand will grow to 175 ZB or 1.75×10^{14} GB by 2025 (in this review, 'B' refers to Byte while 'b' refers to base pair) [3]. With the current storage media having a maximal density of 10^3 GB/mm^3 [4], this will far exceed the storage capacity of any currently available storage method. Meanwhile, the costs of maintaining and transferring data, as well as limited lifespans and significant data losses, also call for novel solutions for information storage [5,6].

On the other hand, since the very beginning of life on Earth, nature has solved this problem in its own way: it stores the information that defines the organism in unique orders of four bases (A, T, C, G) located in tiny molecules called deoxyribonucleic acid (DNA), and this way of storing information has continued for 3 billion years. DNA molecules as information carriers have many advantages over traditional storage media. Its high storage density, potentially low maintenance cost and other excellent characteristics make it an ideal alternative for information storage, and it is expected to provide wide practicality in the future [7].

OVERVIEW OF DNA STORAGE

Research history

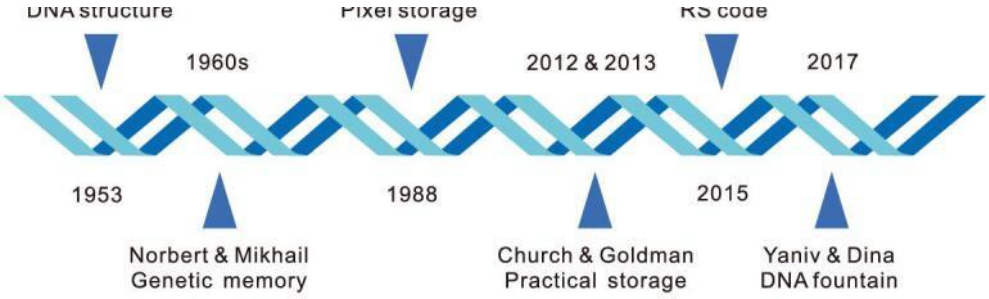
In 1953, Watson and Crick published one of the most fundamental articles in the history of biology in *Nature*, revealing the structure of DNA molecules as the carrier of genetic information [8]. Since then, it has been recognized that the genetic information of an organism is stored in the linear sequence of the four bases in DNA. In just a decade, many researchers had proposed the concept of storing specific information in DNA [9–11]. However, the concept failed to materialize because the techniques for synthesizing and sequencing DNA were still in their infancy.

In 1988, the artist Joe Davis made the first attempt to construct real DNA storage [12]. He converted the pixel information of the image ‘Microvenus’ into a 0–1 sequence arranged in a 5 × 7 matrix, where 1 indicated a dark pixel and 0 indicated a bright one. This information was then encoded into a 28-base-pair (bp) long DNA molecule and inserted into *Escherichia coli*. After retrieval by DNA sequencing, the original image was successfully restored. In 1999, Clelland proposed using a method based on ‘DNA micro-dots’ like steganography to store information in DNA molecules [13]. Two years later, Bancroft proposed using DNA bases to directly encode English letters, in a way similar to encoding amino acid sequences in DNA [14].

However, these early attempts only stored less than tens of Bytes—a small amount of data with little scalability for practical usages. It was not until the first 10 years of the twenty-first century that the groundbreaking work of Church and Goldman led to the return of DNA storage to mainstream interest [15,16]. Church *et al.* successfully stored up to 659 KB of data in DNA molecules, while the maximal amount of stored data before this work was less than 1 KB [17]. Goldman *et al.* stored even more data, reaching 739 KB. It is worth noting that the data stored in the two studies contained not only texts, but also images, sounds, PDFs, etc., which confirmed that DNA can store a wide variety of data types.

Church and Goldman's work led to a research fever of large-scale DNA storage. With increasingly complex compilation methods, the amounts of stored data gradually increased. By the end of 2018, the maximal amount of data stored in DNA exceeded 200 MB, which was stored in more than 13 million oligonucleotides [18]. Along with the development of DNA synthesis and sequencing technologies, new DNA storage methods keep emerging, bringing DNA storage ever closer to practical applications (Fig. 1).

Figure 1.



The history of DNA storage. The figure shows seminal publications in the history of research on DNA storage [8,12,15,16,19,20].

Self-information of DNA molecules

The capacity of a medium to store information is usually measured by the Shannon information. Since the DNA molecule is a heterogeneous polymer composed of a linear chain of deoxyribonucleotide monomers each adopting one of four bases A, T, C and G, the specific arrangement (i.e. sequence) provides a certain amount of information. According to the definition of Shannon information, the maximal amount of self-information (H) that a single base can hold is

$$H = -\sum_{i}^{A,T,C,G} P(i)\log P(i)$$

$$\log \sum_{i}^{A,T,C,G} P(i)\,\frac{1}{P(i)} = \log 4 = 2 \text{ bit},$$

where $P(i)$ represents the probability of base i to occur at any position, and log represents the base 2 logarithm as the bit (binary unit) is usually used as a measurement of digital information [21]. If and only if the four bases are equally likely to occur, that is, P_i = 1/4, each base pair in the DNA molecule can provide the largest information capacity, i.e. 2 bits. The dependence of self-information on base distributions is given in Table 1, where a is the 'probability distribution deviation', that is, the difference between the frequency at which the base appears and the average frequency of 0.25.

Table 1.

Probability distribution of bases and the corresponding self-information values.

a	P_A, $P_T = 0.25 - a$	P_C, $P_G = 0.25 + a$	I(X) (bit/base)	I(X)/I_{max}(X)*
0	0.25	0.25	2	100%
0.001	0.249	0.251	1.999988	99.999%
0.005	0.245	0.255	1.999711	99.986%
0.01	0.24	0.26	1.998846	99.942%
0.05	0.2	0.3	1.970951	98.548%
0.1	0.15	0.35	1.881291	94.065%
0.15	0.1	0.4	1.721928	86.096%
0.2	0.05	0.45	1.468996	73.450%
0.24	0.01	0.49	1.141441	57.072%

*I_{max}(X) = 2 bit/base.

By converting the 2 bit/base to physical density, we obtain

$$\rho = \frac{2\ \text{bit}}{1\ \text{base} \times 325 \frac{\text{Dalton}}{\text{base}} \times 1.67 \times 10^{-24} \frac{\text{g}}{\text{Dalton}}}$$

$$3.69 \times 10^{21} \frac{\text{bit}}{g} = 4.61 \times 10^{20} \frac{\text{Byte}}{g} \approx 460 \frac{\text{EB}}{g},$$

where ρ represents density, 1 EB = 10^{18} B (in this paper, the data storage unit has a radix of 10^3 instead of 1024) and the remaining unit conversion values are derived from ref. [19].

Additional restrictions on the sequence of DNA molecules will further reduce its Shannon information capacity. For example, Erlich *et al.* estimated a Shannon information capacity of ~1.83 bits per base under intrinsic biochemical constraints and technical limitations of DNA synthesis and sequencing procedures [19].

Mutual information and channel capacity

In addition to the self-information carried by DNA molecules, mutual information between channel inputs and outputs is also an important factor in determining information capacity [21]. Mutual information measures the fidelity with which the channel output Y = {y_j|A, T, C, G} (i.e. the readout of a DNA by sequencing) represents the channel input X = {x_i|A, T, C, G} (i.e. the preset DNA sequence):

$$I(X;Y) = \sum_{i}^{A,T,C,G} \sum_{j}^{A,T,C,G} P(x_i y_j)\, I(x_i\, y_j)$$

$$\sum_{i}^{A,T,C,G} \sum_{j}^{A,T,C,G} P(x_i y_j) \log \frac{P(x_i|y_j)}{P(x_i)},$$

and we have

$$I(X;Y) = H(X) - H(X|Y) \leq H(X).$$

For DNA molecules, if each of the four bases corresponds exactly to itself, then H(X|Y) = 0, I(X; Y) = 2 bit/base, and the average mutual information in the transmission is equal to the source entropy, which gives the upper limit of the amount of information transmitted. However, information may be distorted in the process of writing and reading DNA sequences, causing mismatches between the input set X and the output set Y, which reduces the average mutual information during transmission. For example, if each base corresponds to the other three bases except itself with a probability of 1/10, then

$$H\left(Y|x_i\right) = H\left(\frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{7}{10}\right)$$

$$1.35679 \frac{\text{bit}}{\text{base}}.$$

Assuming that the four bases entered are equally probable, we have

$$H(Y|X) = 4 \cdot \frac{1}{4} \cdot H\left(Y|x_i\right) = 1.35679 \frac{\text{bit}}{\text{base}},$$

while

$$H\left(X\right) = H\left(Y\right) = 2 \frac{\text{bit}}{\text{base}}.$$

The joint entropy of X and Y is

$$H(XY) = H(X) + H(Y|X)$$

$$3.35679 \frac{\text{bit}}{\text{base}},$$

and the average mutual information is

$$I(X;Y) = H(X) + H(Y) - H(XY)$$

$$0.64321 \frac{\text{bit}}{\text{base}}.$$

Thus, the distortion of the base readout greatly reduces the utility of information transmission in DNA. Table 2 shows the average mutual information at different transmission error rates m_i (the probability that one base is incorrectly read out as one of the other three bases), assuming 2 bit/base inputs. Figure 2 gives the variation of the average mutual information as a function of the input base bias and the transmission error rate.

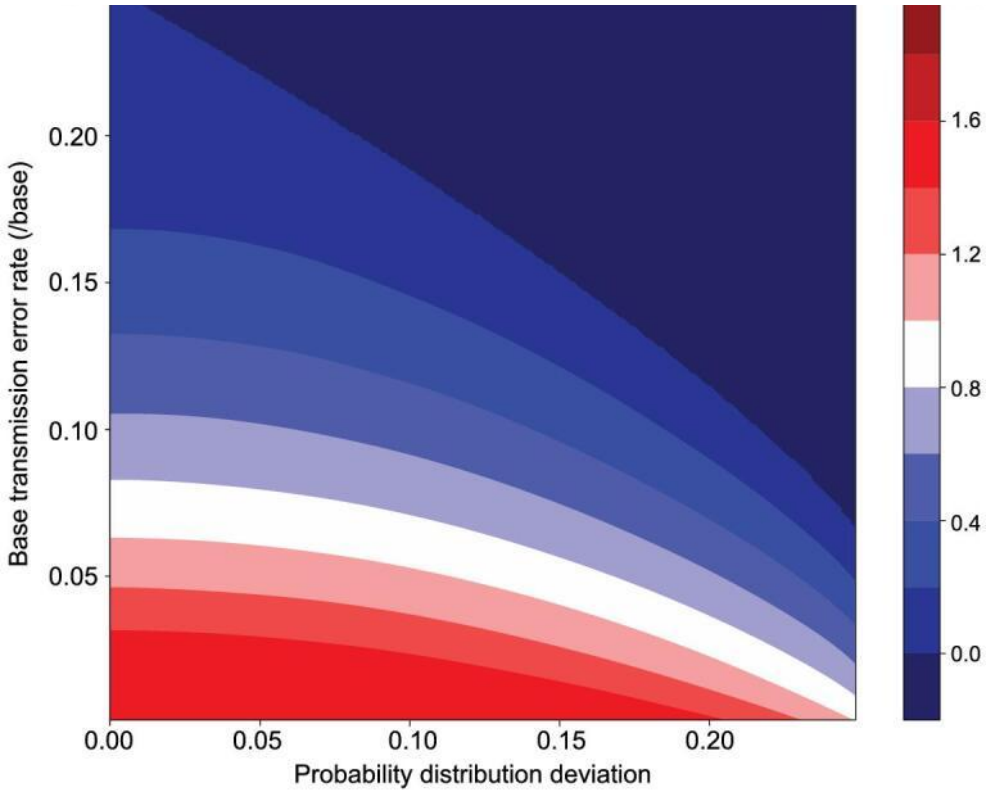
Table 2.

Base transmission error rate and the corresponding mutual information.

mi	**0.001**	**0.005**	**0.01**	**0.02**	**0.05**	**0.1**	**0.2**
I(X;Y) (bit/base)	1.9658	1.8639	1.7581	1.5774	1.1524	0.64321	0.0781
I(X;Y)/I_{max}(X;Y)*	98.29%	93.19%	87.91%	78.87%	57.62%	32.16%	3.95%

*I_{max}(X; Y) = 2 bit/base.

Figure 2.



The relationship among the average mutual information transmitted by DNA, the probability distribution deviation of bases and the base transmission error rate. Color indicates the average mutual information values.

More generally, the channel transmission characteristics of DNA molecules can be defined by a 4 × 4 transfer matrix *T* [21]

$$XT = Y,$$

where *X* is the input set and *Y* is the output set, and *T* can be expanded as:

$$T = \begin{bmatrix} P_{AA} & P_{AT} & P_{AC} & P_{AG} \\ P_{TA} & P_{TT} & P_{TC} & P_{TG} \\ P_{CA} & P_{CT} & P_{CC} & P_{CG} \\ P_{GA} & P_{GT} & P_{GC} & P_{GG} \end{bmatrix},$$

where P_{ij} refers to the probability that the input base i is received as base j after channel transmission. If

$$\begin{cases} P_{ij} = 1, & i = j \\ P_{ij} = 0, & i \neq j \end{cases},$$

the above transfer process corresponds to the passage of information through an ideal channel. In reality, the values of P_{ij} can be obtained for a specific storage method through systematic experimentation. We can therefore obtain

$$H(Y|x_i) = \sum_{j}^{A,T,C,G} H(P_{ij}).$$

If we denote by P_i(i = A, T, C, G) the frequency of each base in a channel input, then the corresponding frequency distribution in the output Y, as well as the average mutual information, is completely determined by P_i and the transfer matrix T

$$P_i' = \sum_{j}^{A,T,C,G} P_j \cdot P_{ji}.$$

Therefore, we can obtain

$$H(Y) = \sum_{i}^{A,T,C,G} H(P_i'), H(Y|X)$$

$$\sum_{i}^{A,T,C,G} P_i \sum_{j=1}^{A,T,C,G} H(P_{ij}),$$

and the average mutual information is

$$I(X;Y) = H(Y) - H(Y|X)$$

$$\sum_{i}^{A,T,C,G} H\left(\sum_{j}^{A,T,C,G} P_j \cdot P_{ji}\right) - \sum_{i}^{A,T,C,G} P_i \sum_{j=1}^{A,T,C,G} H(P_{ij}).$$

Due to the non-negative nature of the entropy function, the average mutual information can only be maximized when the latter term is 0. This requires that all P_{ij} values be either 0 or 1, i.e. X and Y form a strict one-to-one mapping relationship. It is not necessary for each base to correspond to itself, though. For example, if all A in the DNA molecule become T after channel transmission and T→C, C→G, G→A, the maximal mutual information can also be achieved. In practice, this method is cumbersome and

unnecessary. However, this approach may have potential uses in information encryption [22].

For a specific storage method with its measured transfer matrix *T*, one may find the input base probability distribution that generates the highest channel capacity [23]

$$C = \max_{P(X)} \{I(X;Y)\},$$

which requires

$$\frac{\partial I(X;Y)}{\partial P_i} = 0.$$

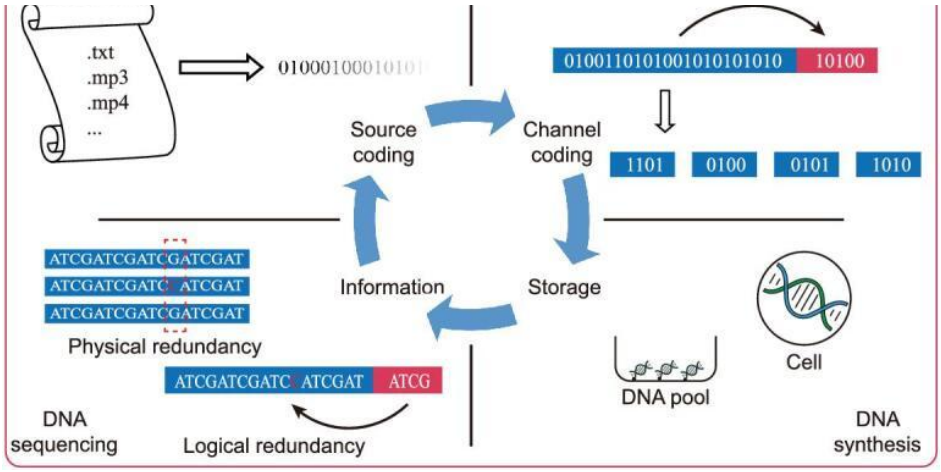
After substituting the previously obtained expression for *I*(X; Y), the best input probability distribution can be obtained by calculation.

In addition to mismatches, common errors in synthesis and sequencing include insertions and deletions, collectively called indels. Generally, the impact of indels on information storage is much greater than that of mismatches, since the loss or gain of consecutive sequences may nullify the entire DNA molecule. In next-generation sequencing such as Illumina, indels occur less than 1% as frequently as substitutions do. However, single-molecule sequencing has been reported to be prone to indel errors [24]. Indels in DNA storage correspond to 'erasure channels' in the field of information science. Theory on this subject is still under active development. Various models of erasure channels have been established. We refer the readers elsewhere (e.g. ref [25]) without elaboration here.

IMPLEMENTATION OF DNA STORAGE

Figure 3 summarizes the general workflow of the DNA information storage process.

Figure 3.



Flow of information in DNA-based information storage. Top left: source coding, i.e. converting information into binary code (or other radix) series. Top right: channel coding, i.e. data error detection/correction coding, providing an error correction/error detection capability by providing additional bits of redundancy. Bottom right: information storage. After the desired DNA molecule is synthesized, it can be stored *in vitro* or *in vivo*. Bottom left: information readout. Each part will be detailed in the text.

Source coding

In order to use DNA molecules for information storage, information must first be converted into a sequence of four bases in the DNA molecule. In general, each base is equivalent to a quaternary number, corresponding to two binary digits. Obviously, any digital information can be encoded into the DNA molecule by a simple conversion. This applies to all types of data that can be stored on a hard drive.

In the field of information science, different data types are processed using different encoding and compression algorithms [23]. Here, we take the classic text-file format as an example to introduce the various compilation methods of DNA storage. In the first attempt by Bancroft *et al.*, English letters were directly encoded by base triplets in a manner like the amino acid codon table, for example, 'AAA' represents the letter 'A' [14]. Interestingly, they only used three bases to form a 'ternary digit', while G was reserved for sequencing primers. The method ignored capitalization because three bases can produce a coding space of only $3^3 = 27$ elements, which is just enough to encode 26 letters. And, by the same reason, this encoding scheme does not apply to other data types.

A pioneering study by Church *et al.*, as the first big volumne DNA storage work, used a more scalable approach. They first converted different files into binary sequences in the *HTML* format and then converted these into DNA sequences [15]. In comparison, Goldman *et al.* applied the Huffman coding scheme in the first step, which employs ternary instead of binary conversion. Huffman coding simultaneously compresses the

data and this is the first DNA storage study in which data compression algorithms were used.

In fact, data compression is essential when scaling DNA storage to larger data volumes. For text files, many lossless data compression algorithms exist that greatly reduce the space required to store them. The lower bound of the storage space in a lossless compression scheme is defined by Shannon's first theorem. If the source entropy of a discrete memoryless stationary source is $H(X)$, using the r-ary symbol to encode the N-time extended source symbol sequence of the source in variable length, there must be a unique distortion-free and decodable code [21], with the average code length L satisfying

$$\frac{H(X)}{\log r} \leq \frac{L}{N} < \frac{H(X)}{\log r} + \frac{1}{N}.$$

The text files currently stored in the DNA molecules are treated as memoryless sources (i.e. there is no correlation between adjacent letters, $N = 1$). When binary encoding ($r = 2$) is used, the average code length L satisfies

$$H(X) \leq L < H(X) + 1.$$

Intuitively, the average code length of each symbol in the code cannot be less than the source entropy

$$H = -\sum_i p(i) \log p(i),$$

where i represents each letter in the text file and $p(i)$ is the frequency at which it appears. The available algorithms for text compression include Huffman coding, arithmetic coding, dictionary coding, etc., among which Huffman coding is the most commonly used in the field of DNA storage. This is a variable-length code that uses shorter codes for high frequency letters and longer codes for low frequency letters to

reduce the average code length of the text file. The Huffman coding algorithm is readily applicable to any text file and is compatible with special characters.

It is worth mentioning that, for a particular language, it is possible to encode a piece of text with a shorter code length. In English, for example, the frequency of the 26 letters in a typical text varies greatly. If they are assumed to be statistically independent, they are equivalent to a discrete memoryless source. Statistical analyses revealed the average source entropy of English texts is [26]

$$H_s = -\sum_{i=1}^{27} p(i) \log p(i) = 4.02 \left(\frac{\text{bit}}{\text{letter}} \right).$$

However, in a text context, English letters are in fact not statistically independent. Shannon studied the English text as an n^{th}-order Markov source. For $n \to \infty$, he obtained the statistical inference value [26]

$$H_\infty = 1.4 \left(\frac{\text{bit}}{\text{letter}} \right).$$

H_∞ is called the limit entropy. For any finite n, it is possible to compress information to reach density H_n($H_\infty < H_n < H_s$) by considering the context dependencies among letters.

Channel coding

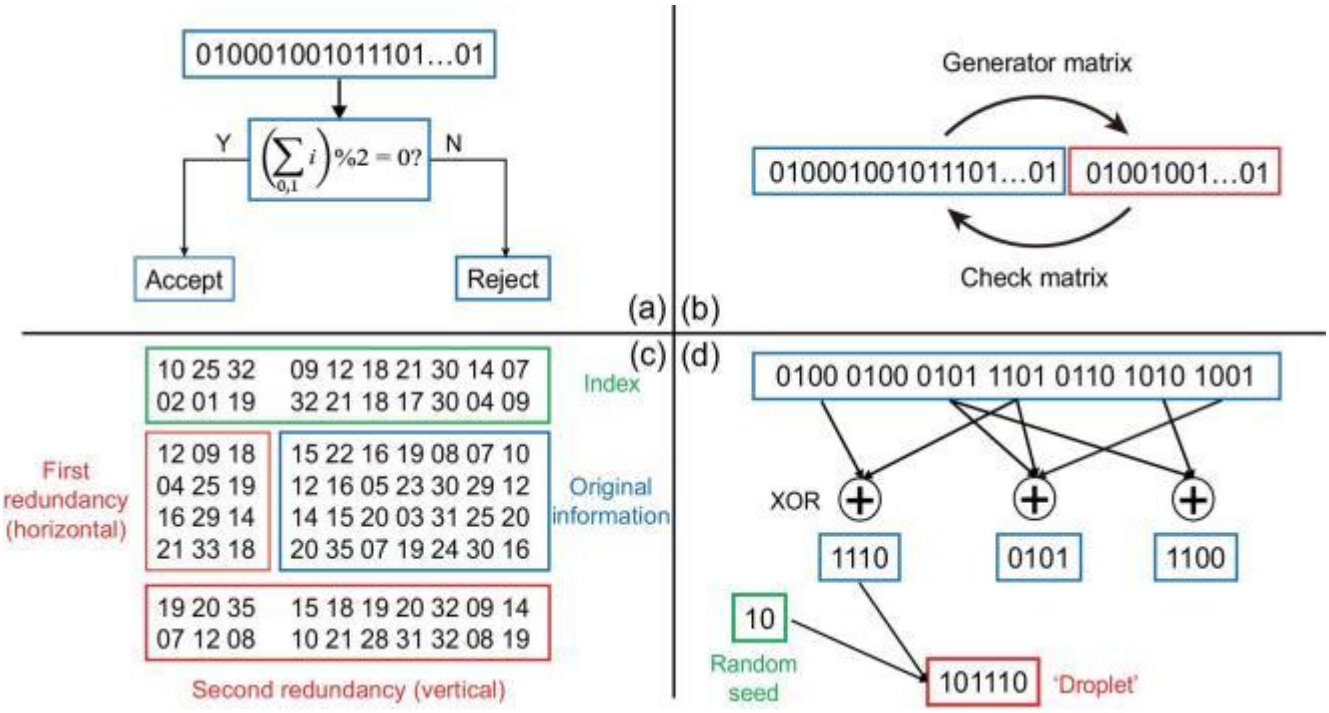
Information distortion often occurs during transmission [21]. For DNA molecules, errors may occur during synthesis, replication and sequencing. There are two ways to recover raw data despite information distortion: physical redundancy and logical redundancy. Physical redundancy entails increasing the copy number of DNA molecules that encode the same information. For example, Goldman *et al.* used 4-fold redundant DNA molecules to store information in their initial attempts, i.e. in each short DNA molecule of 100 bp long, the first 75 bp overlapped with the previous molecule and the last 75 bp overlapped with the next molecule [16]. Previous work by Nozomu *et al.* used different

sequences to encode the same information. In the process of mapping the binary 0–1 sequence to DNA bases, a binary number was shifted each time and the corresponding base sequences were obtained. As a result, they were able to encode the same information using four different base sequences [27].

Sequencing coverage also contributes to physical redundancy. In the initial work of Church *et al.*, the sequencing coverage was 3000× [15]. However, physical redundancy is not sufficient for achieving lossless data transmission. The work of Goldman and Church failed to completely restore all the information. Church *et al.* found a total of 22 errors in the sequencing results [15] and Goldman *et al.* also obtained sequences that cannot be automatically recovered [16]. In addition, for large data volumes, physical redundancy imposes a dramatic increase in costs.

Another way to correct errors is by logical redundancy—a method widely used in the communication field. The general idea of logical redundancy is to add extra symbols, called ‘check symbols’ or ‘supervised symbols’, in addition to the symbols encoding information. When the information symbols are incorrect, the check symbols can be used to detect or correct errors so that the information can be accurately recovered (Fig. 4).

Figure 4.



Illustrations of channel coding for DNA storage. (a) Hamming code, which can only be used to check one error. (b) Linear block code. (c) RS code. Shown here is the two-round RS code used by Grass *et al.* [20]. (d) Fountain code. Shown here is the LT code used by Erlich *et al.* [19].

The most commonly used error correction code is the linear block code (Fig. 4b). Specifically, if a group of information symbols has a length of k, a check symbol of length r can be added using a specific generator matrix to obtain a linear block code with a code length of $n = k + r$. Once the generator matrix is selected for a set of codes, the pairing between the information symbols and the check symbols determines whether a codeword is legal or not. The apparent coding efficiency of this code is k/n and the error correction capability scales with $r/n = 1 - k/n$. Thus, there is a trade-off between the coding efficiency and the error correction capability.

The most basic class of linear block codes is the Hamming code (Fig. 4a). Simple as it is, only one error can be detected in each group of code words. Due to its obvious limitations, the Hamming code has not been used for DNA storage. Another class of linear block code is called the cyclic code, by which each group of codewords is still legal after one cyclic shift. The most widely used type of cyclic code is the Bose–Chaudhuri–Hocquenghem (BCH) code, which is a code class that can correct multiple random errors based on the Galois binary field and its extension. To avoid crossover between the information symbol and the check symbol, one can use a generator polynomial to get a special BCH code, which is called a system code [21].

Quantitative assessments can be performed to compare the usefulness of physical redundancy and logical redundancy. For second-generation sequencing, several studies on DNA storage in recent years have pointed out that the total error rate in the synthesis–storage–sequencing process (equivalent to channel transmission) is about 1% [28,29]. Assuming misread events are independent and identically distributed, their total number follows the Poisson distribution. For instance, for a DNA molecule of 128 bp in length, the probability of any error occurring is

$$P_E = 1 - \frac{e^{-128\times\ 0.01} \times (128 \times 0.01)^0}{0!} = 0.722.$$

If 3-fold physical redundancy is used for error correction, the error becomes incorrectible when more than two of the three copies are misread for the same base at the same site. Therefore, the probability of an uncorrected error is

$$\sum_{i=0}^{1} C_3^i \cdot 0.99^i \cdot \left(1-0.99\right)^{3-i} = 0.000298,$$

at any base. For the 128-bp DNA molecule, the probability of any error occurring is

$$P_E = 1 - \frac{e^{-128\times 0.000298} \times 0.0128^0}{0!}$$

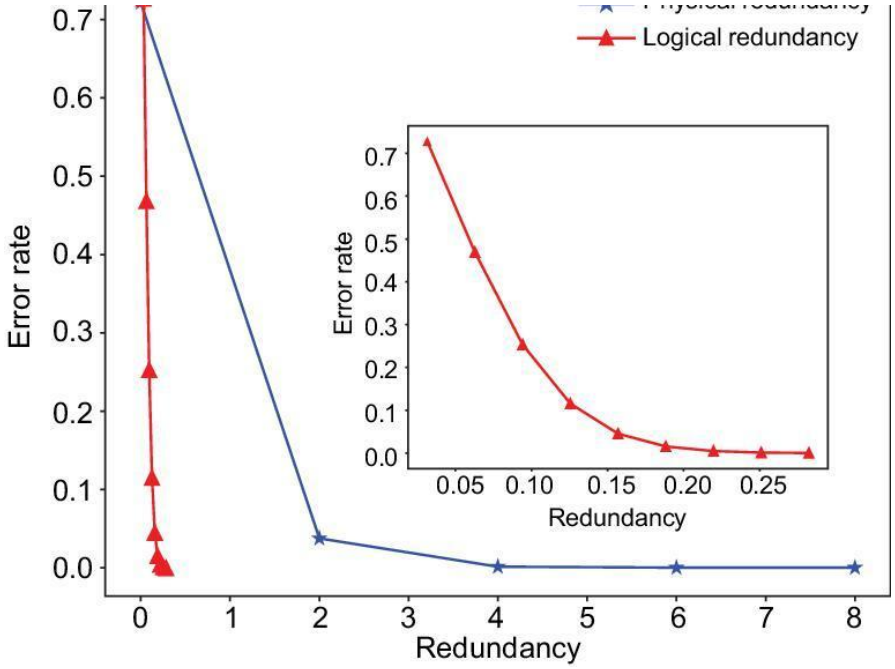
$$0.03742568.$$

Now let us turn to logical redundancy. We will use the (255, 207) BCH code as an example (note that this corresponds to the above 128-bp DNA molecule), which can correct six errors in each group of 255-bit symbols. Still using the overall error rate of 1% per base, the code fails to correct all errors only when at least seven errors occur in a group of code words, which has a probability

$$P_E = 1 - \sum_{i=0}^{6} \frac{e^{-2.55} \times 2.55^i}{i!} = 0.016.$$

It can be seen that a logical redundancy of <20% already suppresses error rates to a similar extent as a physical redundancy of 200% does. Shannon's second theorem states that, for a discrete memoryless channel with capacity *C* and a discrete source with entropy per second *R*, if $R \leq C$, then, as long as the code length *n* is large enough, an encoding rule and a corresponding decoding rule can always be established to make the average error probability P_E arbitrarily small. Figure 5 compares varying degrees of physical and logical redundancy and their error-correction capabilities.

Figure 5.



The error correction capacity of coding systems with different levels of physical and logical redundancies. The 'error rate' on the *y*-axis refers to the probability of not being able to correct all the errors. Blue line: the effect of physical redundancy on error correction capacity (taking 128-bp DNA as an example). Red line: the effect of logical redundancy on error correction capacity. Here, an original BCH code with code length $n = 255$ is used as an example. Inset: magnified view of logical redundancy.

The Reed-Solomon (RS) code that has been applied in DNA storage is a special non-binary BCH code, which has been widely used in fiber, satellite and deep-sea communication, etc. [21]. Grass *et al.* used the RS codes generated on the Galois Field GF (47) for error correction [20]. Notably, they added two rounds of RS codes, called the 'inner code' and the 'outer code', respectively, to map the information symbols along orthogonal directions (Fig. 4c). The outer code also mapped the indices. This type of coding is optimized to correct bursts of errors, such as in the case of consecutive base losses, i.e. sequence degradation. In addition, RS codes were included in the 'DNA fountain' system used by Erlich *et al.*, where they were not used for error correction, but for detecting and discarding erroneous sequences [19].

By contrast, fountain coding uses a completely different framework than linear block codes, amounting to a codeless erasure code. The basic idea is to group the signal sources into smaller packets. After obtaining an adequate number of packets, the original information can be successfully restored (Fig. 4d). The main advantage of the fountain code is its extremely low redundancy and it can handle 'erase' (deletion and insertion of bases) errors. Erlich *et al.* used the classic Luby Transform Code in the fountain code, i.e. the LT code. If DNA molecules are lost to varying degrees, the LT code can still handle it well through detailed design. Currently, the fountain code may be

the only error-correction code in the field of DNA storage that can robustly deal with the loss of DNA molecules. The success of commercial LT codes for digital information (achieving a decoding failure rate $<10^{-8}$ with <5% redundancy [30]) has highlighted its potential for DNA storage.

Encoding information in DNA sequences

After being converted to a binary (or other radix) sequence, the information needs to be transformed into base sequences in DNA. For binary data, the most intuitive conversion is representing 2 bits with one base. The correspondence can be set arbitrarily to control the base compositions in a specific DNA molecule. Furthermore, this method provides the maximal information storage capacity. However, it may result in sequences that are difficult to manipulate, such as long tracts of homopolynucleotides that are error-prone in high-throughput sequencing [31].

Much of the previous work was focused on solutions to this problem. Church *et al.* used one base to represent a single binary digit (i.e. A, C = 0; G, T = 1), so that alternative bases can be adopted to avoid homopolynucleotide tracts [15]. However, the low information density prevented its use in later studies. Goldman *et al.* pioneered in a ternary base conversion table that allows each base to represent a ternary number depending on the previous base [16]. This approach absolutely avoids homopolynucleotide tracts without compromising information density. In the fountain coding scheme by Erlich *et al.*, a single base can still correspond to two binary digits, with unqualified sequences discarded altogether in transmission [19]. They further analysed the constraint on the GC content of DNA molecules as it affects the stability of DNA molecules, the substitution and indel error rates during sequencing, and the dropout rates in PCR amplifications, which were also emphasized in other work [32]. An appropriate GC content close to 50% can be obtained through proper base encoding methods as well as by sequence screening—that is, selecting DNA molecules with appropriate GC ratios to store information while discarding molecules with unreasonable GC contents. In the sequence screening scheme, Erlich *et al.* gave an estimate of 1.98 bits/nt for the maximal coding capacity of DNA storage considering the effects of homopolymers and GC contents, although the latter contributes a comparatively small reduction [19].

Information density of DNA storage

As shown in the previous section, the upper limit of the information storage density of DNA has been calculated to be about 4.606×10^{20} Bytes/g, but a more practical indicator is the volumetric density. In the initial work of Church *et al.* [15], the bulk density of DNA molecules was approximated to the density of pure water, which gave an information density of 4.606×10^{17} Bytes/mm^3. In comparison, the information storage density of classic media, such as flash drives, optical tape and hard disks, is of the order of 10^9 Bytes/mm^3 [4,5].

However, the estimate was made under ‘ideal conditions’, ignoring many practical factors. First, the theoretical bulk density can hardly be reached, as DNA molecules need to be stored in specific environments to prevent degradation. For example, most *in vitro* DNA storage studies were based on short DNA oligonucleotides (oligos) in a DNA pool, which was dissolved in dilute solution. Second, physical and logical redundancies reduce the actual information density to various extents. Third, a certain length of index is needed in the DNA molecules to provide addresses, which are themselves not available for storing information.

Here, we briefly analyse the indexing demand of *in vitro* DNA oligo storage. Due to technical bottlenecks in the current DNA synthesis process, most studies to date have used 150- to 250-bp oligos as storage units. Since DNA oligos are fully mixed in a library, a unique index needs to be assigned to each oligo encoding unique information. Table 3 shows the length of the index required in a 200-bp molecule when storing different amounts of data. When the length of the index in this sequence is k bp, the number of indexable molecules is 4^k and the number of bits used to store information is $400–2k$ per molecule. Therefore, the total storage capacity of the oligo pool is

$$Q(k) = (400 - 2k) \cdot 4^k \text{ bit.}$$

Table 3.

Index length required to store different amounts of data with 200-bp DNA molecules.

Data amount	1 Byte	1 KB	1 MB	1 GB	1 TB	1 PB	1 EB	1 ZB
Index length (bp)	0	3	8	13	18	23	28	33
Index ratio*	0	1.5%	4%	6.5%	9%	11.5%	14%	16.5%

*Index ratio = *L*(index)/*L*(molecule).

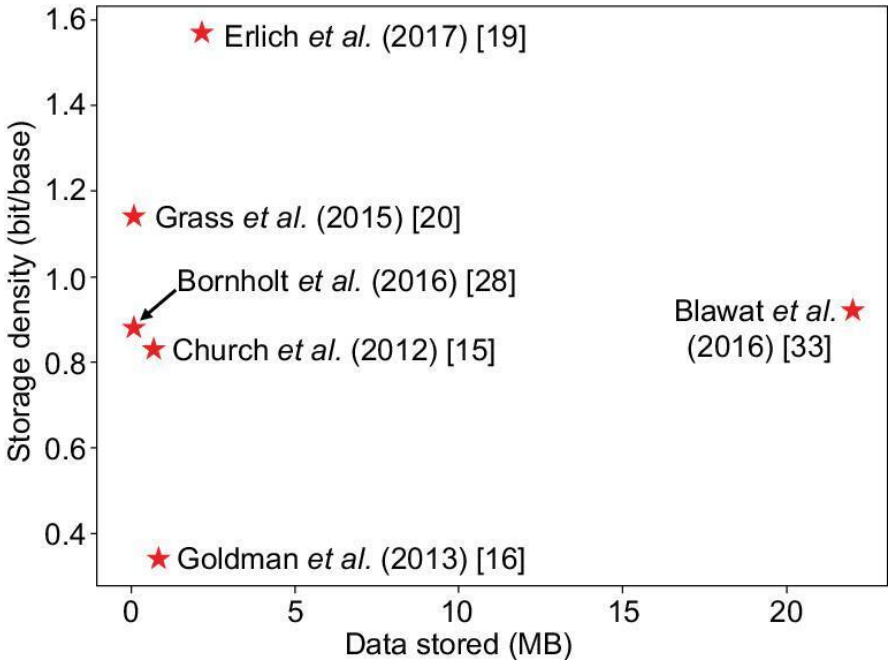
In reality, it is almost impossible to store ZB orders of data in a single DNA oligo library. For example, the dilute solution condition, as is required for efficient information retrieval and amplification, is hardly met, with $4^{33} \approx 10^{20}$ molecules dissolved in a few liters of solution. Another constraint is imposed by the free diffusion of DNA oligos in solution. Although, in the 100 base pair range, the diffusion coefficient of DNA oligos can be higher than 10 $\mu m^2/s$, the Brownian motion of oligos could not traverse a significant portion of the reaction system in a reasonable reaction time to enable searching of the probes for random access of information, especially in large libraries. Our crude calculations suggest an upper limit of PB information in a 1-liter reaction system. Lastly, the theoretical indexing limit should not be saturated to ensure sufficient specificity of indices against probes. One possible solution for the storage of large data volumes is to use physically separated DNA pools. This has not been explored yet, due to the

extremely limited amount of information that has been stored in DNA so far. However, as DNA storage comes close to real practice, rigorious systems design such as this will be needed.

Finally, as mentioned in the previous sections, intrinsic limits of DNA synthesis and sequencing technologies impose constraints on the DNA sequences that could code information reliably, which reduces the information storage density of DNA molecules (e.g. Fig. 2).

Figure 6 shows the amounts of data stored and the data storage densities achieved in major DNA data storage publications since 2012.

Figure 6.



Amounts of data stored and storage densities achieved in major DNA data storage studies. The storage density refers to the effective density, i.e. the total amount of information stored divided by the total number of bases used (number of oligonucleotides × number of bases per oligonucleotide molecule). The *x*-axis shows the total amount of data stored [15,16,19,20,28,33].

TECHNICAL ASPECTS AND PRACTICAL CONSIDERATIONS

DNA synthesis and assembly technology

The past few decades have witnessed the rapid development of DNA synthesis and assembly technologies, which laid the groundwork for the advancement of novel fields and technologies including DNA information storage.

The first generation of DNA synthesis techniques are based on solid-phase phosphoramidite chemistry [34,35]. The main advantage of this method is its high accuracy, albeit with a high cost and a low throughput. Moreover, for the consideration of sequence integrity and synthesis efficiency, the product length is limited to 150–200 bp. The second-generation, array-based DNA synthesis is a technique for synthesizing DNA using a series of electrochemical techniques on microarray chips. In each cycle, nucleotides are conjugated to DNA strands at specific locations of the chip, allowing simultaneous elongation of a heterogeneous pool of oligos [36]. Array-based DNA synthesis significantly improved the speed, efficiency and cost-effectiveness of DNA synthesis. In particular, the 10^6 parallel throughput achieved on current state-of-the-art second-generation platforms increases the total speed of synthesis to a few kilo bases per second. The third-generation DNA synthesis techniques are based on enzymatic synthesis. Although still in their infancy, they are expected to dramatically reduce the time and cost of DNA synthesis. Lee *et al.* gave an estimate of 40 s/cycle, which is six times as fast as phosphoramidite synthesis, and a projected reduction in cost by several orders of magnitudes once their terminal deoxynucleotidyl transferase (TdT) enzymatic reaction system is miniaturized [37].

In addition to DNA synthesis technology, DNA ligation and assembly technologies will provide powerful support for DNA information storage and in particular long-chain DNA storage. At present, commonly used DNA amplification, ligation and assembly techniques include PCR [38], loop-mediated isothermal amplification (LAMP) [39], overlap-extension PCR (OE-PCR) [40], circular polymerase extension cloning (CPEC) [41], InFusion technology [42], sequence- and ligation-independent cloning (SLIC) [43], restriction enzyme digestion and ligation [44], as well as Gibson [45] and Golden Gate assembly [46–48].

DNA sequencing technology

Since the invention of the Sanger sequencing method in 1977, DNA sequencing has developed into a fully fledged technology, with its cost dropping by 100 000 times in recent years [49]. Based on the underlying mechanisms, DNA sequencing is generally divided into three generations: Sanger sequencing, high-throughput sequencing/Next Generation Sequencing (NGS) and single-molecule sequencing.

The first generation of sequencing technology is based on Sanger's double-deoxygenation termination sequencing combined with fluorescent labeling and capillary array electrophoresis [50]. Currently, automated first-generation DNA sequencing is still widely used.

The core idea of NGS is large-scale parallel sequencing, which enables the simultaneous sequencing of hundreds of thousands to millions of DNA molecules with short read lengths. The available platforms include Roche/454 FLX, Illumina/Solexa Genome Analyzer, HiSeq and ABI/Applied Biosystems SOLID system, Life

Technologies/Ion Torrent semiconductor sequencing, etc. [51–54]. NGS has raised the sequencing throughput from 100 Kb to the orders of Gb and Tb, and reduced the cost of sequencing at a rate four times that predicted by Moore's Law [49].

The Helicos/HeliScope single-molecule sequencer [55], Pacific Biosciences SMRT technology [56,57], Oxford Nanopore Technologies nanopore single-molecule technology [58,59] and single-cell genomic sequencing technology [60] are considered third-generation single-molecule sequencing technologies. Besides removing the dependence on PCR amplification, third-generation sequencing has managed to significantly increase the read length and raise the read speed. The cost and accuracy are currently less than satisfactory but are expected to improve with further technological development, making it more practical for the purpose of DNA information storage [52–60]. Table 4 compares performance of typical sequencing techniques from the three generations.

Table 4.
Comparison of three generations of DNA sequencing technology [50,52–54].

Sequencing technology	First generation (Sanger)	Second generation (Illumina)	Third generation (ONT nanopore)
Cost (per Kb)	$1–2	$$10^{-5}$$–$$10^{-3}$$	$$10^{-4}$$–$$10^{-3}$$
Error rate	0.001–0.01%	0.1–1%	~10%
Sequencing length	1 Kb	25–150 bp	200 Kb
Read speed (per Kb)	~$$10^{-1}$$ h	~$$10^{-7}$$–$$10^{-4}$$ h	$$10^{-7}$$–$$10^{-6}$$ h
Sequencing throughput	1 Kb	$$10^{8}$$–$$10^{12}$$ bp	$$10^{9}$$–$$10^{13}$$ bp

Cost of DNA data storage

Compared to traditional data storage methods, DNA storage has significantly lower storage maintenance costs. For example, if a data center stores $$10^{9}$$ G data on tape, it will require as much as $1 billion and hundreds of millions of kilowatts of electricity to build and maintain for 10 years [5]. DNA storage can reduce all these expenses by 3 orders of magnitude [5]. Nevertheless, the cost of DNA synthesis can be significant and it will become a limiting factor for DNA storage to commercialize. At the current cost of ~$$10^{-4}$$/base [61] and a coding density of 1 bit/base, a conservative estimate of the write cost is $800 million/TB, while tape costs about $16/TB [62]. On the other hand, the read cost achieved by current sequencing technologies is orders of magnitude smaller, at ~$0.01–1 million/TB [63]. However, it is expected that the cost of DNA synthesis and sequencing will continue to decrease in the future, and new techniques and methods will be applied to DNA storage [52].

The age limit of DNA storage

DNA molecules naturally decay with a characteristic half-life [64,65], leading to a gradual loss of stored information. The half-life of DNA highly correlates with temperature and the fragment length. For example, Allentoft concluded that a DNA molecule of 500 bp has a half-life of 30 years at 25°C, which extends to 500 years for a fragment of 30 bp. Interestingly, fossils provide empirical evidence of DNA's stability over thousands of years [65]. In this case, stability is significantly improved by low temperatures and waterproof environments [65]. Indeed, at −5°C, the half-life of the 30-bp mitochondrial DNA fragment in bone is predicted to be 158 000 years [65]. Some studies have suggested that DNA can be placed in the extremely cold regions of Earth or even on Mars for millennium-long storage. Other studies have explored packaging materials for DNA molecules and have demonstrated impressive stability [66,67]. Grass *et al.* encapsulated solid-state DNA molecules in silica and showed that they had better retention characteristics than pure solid-state DNA and DNA in liquid environments [20]. Judging by first-order degradation kinetics, they concluded that it could survive for 2000 years at 9.4°C or 2 million years at −18°C, surpassing all potential quantitative data storage materials invented to date. It is reasonable to expect a long lifetime for data stored in DNA even at room temperature, which makes DNA storage especially suited for cold data with infrequent access. Further research may extend the lifetime of DNA storage over the duration of human civilization with minimal maintenance.

In vivo DNA storage

Most DNA storage attempts to date were done *in vitro*. However, the genomic DNA of living cells has become an ideal medium for information storage due to its durability and bio-functional compatibility. Its advantages are becoming more obvious with the improvement of throughput and reduction in cost of DNA synthesis and sequencing technology [15,16,19]. Compared to *in vitro* DNA storage, *in vivo* storage takes advantage of the efficient cellular machineries of DNA replication, proofreading and long-chain DNA maintenance, offers the chance for assembly-free random access of data [18], and supports live recording of biochemical events *in situ* in living organisms as a generalized concept of information storage.

The development of synthetic biology and gene editing technologies have allowed us to change genetic information with unforeseen flexibility and accuracy [68,69]. Natural and engineered DNA targeting and modifying enzymes can be used as write modules in DNA storage systems, and the toolbox of DNA writers is rapidly expanding and improving in terms of programmability and accuracy [68–73]. The work of Shipman *et al.* offers an example for large-scale *in vivo* DNA storage. A library of indexed short DNA fragments encoding 2.6 KB of information was distributively inserted into the CRISPR arrays of multiple live bacterial genomes in a heterogenous population. For complete information retrieval, DNA from different cells was collected and sequenced, and the original information is reconstructed by proper alignment [74]. Yang *et al.* stored a total of 1.375 Bytes of information in the *E. coli* genome by different integrase enzymes [75]. Bonnet *et al.* used recombinases to write and erase information in living cells [76].

DNA writers can be broadly categorized into precise and pseudorandom writers on the basis of the mutational outcomes [68]. Precise DNA writers, including site-specific recombinases [72], reverse transcriptases [77] and base editors [78], generate predetermined mutations, whereas pseudorandom DNA writers, including site-specific nucleases [79–81] and the Cas1–Cas2 complex [79], generate targeted but stochastic mutations.

Site-specific recombinases are a class of highly efficient and accurate DNA writers that can flip, insert or excise a piece of DNA between their cognate recognition sites. Using recombinases, the information is heritably stored in a specific genomic location [72,75,80]. On top of this, reversible writing of information can be achieved by adding another enzyme (the excisionase), which erases the previously written information and resets the state of DNA [76]. The second class of precise DNA writers relies on reverse transcriptases [68,77]. For example, the SCRIBE (Synthetic Cellular Recorders Integrating Biological Events) system is activated in response to a specific stimulus (such as a chemical), producing a programable DNA sequence change [82]. The third class performs nucleotide-resolution manipulation of DNA via base editing [68,78], such as CAMERA (CRISPR-mediated analog multi-event recording apparatus) [83], generating deoxycytidine (dC)-to-deoxythymidine (dT) or deoxyadenine (dA)-to-deoxyguanine (dG) mutations.

Pseudorandom DNA writers relies on targeted double-stranded DNA breaks generated by site-specific nucleases [68], including Cas9, ZFNs and TALENs [79–81]. However, the write efficiency is highly dependent on the nonhomologous end joining pathway, which is lacking in many model organisms [79–81]. A second class of pseudorandom DNA writers leverages the cellular immune functionality of the Cas1–Cas2 system, which integrates information-encoding short ssDNA fragments (approximately 20–30 bp) into the CRISPR array in an oriented fashion [84].

For *in vivo* DNA storage, it is essential to consider the maximal amount of information that a single cell can carry. At present, *E. coli* is the most thoroughly studied prokaryote, but other microorganisms might be used for DNA storage as well. In an interesting example, Mitsuhiro *et al*. cloned the 3.5-Mb genome of the photosynthetic bacterium *Synechocystis* PCC6803 (3.5 Mb) into the 4.2-Mb genome of *Bacillussubtilis* 168, producing a 7.7-Mb chimeric genome [85]. This suggests a surprisingly large tolerance of prokaryotic cells in foreign DNA. If a cell can hold 4 Mb of DNA, it is possible to store 8 Mbit, or 1 MB, of information. In this scenario, a homologous recombination system handling long DNA fragments works more efficiently than a CRISPR-based system dealing with short fragments.

However, incompatibility and interactions between the information-carrying DNA and the host DNA pose challenges for *in vivo* DNA storage. For example, when Mitsuhiro *et al*. attempted to insert the exogenous genome into the genome of *B. subtilis*, efficiency was significantly affected by the host genome's symmetry [85]. As far as biosafety is

concerned, although artificially encoded DNA is not prone to forming open reading frames, misexpression may emerge as the storage volume rises, and its biological consequences should be subject to close scrutiny. On the other hand, there is not enough evidence to show whether the insertion of DNA fragments affects the host cell's own gene expression. In eukaryotic cells, the problem is further complicated by the presence of a wide range of *cis*-acting elements. Effective methods must be devised to prevent the potential biological impacts associated with the insertion of DNA fragments carrying non-biological information.

THE FUTURE OF DNA STORAGE

Prospects and challenges

Although DNA information storage has enormous application potential, many problems need to be addressed before its broader implementation. First, the cost of writing and reading information is still prohibitively high and the efficiency of storing data is too low. However, DNA synthesis and sequencing costs have been reduced by 10-million-fold over the past 30 years, and the trend will continue to meet the needs of practical DNA storage in the foreseeable future [49,51]. It is predicted by the Molecular Information Storage Program that DNA synthesis cost will reduce to $\$10^{-10}$/bp by 2023 [86]. At the same time, the read and write speeds have gradually increased. In their original study (2012), Church *et al.* concluded that DNA synthesis and sequencing technologies require improvements of 7–8 and 6 orders of magnitude, respectively, to compete with current information read and write speeds [15]. The data presented by Goldman *et al.* show that the main contributor to the cost of DNA storage is synthesis and, based on their calculations, if the cost of synthesis is reduced by another 2 orders of magnitude (compared to 2013), DNA storage will outperform magnetic medium storage for decade-long data storage—a goal that could be achieved in just a few years [16]. In 2017, Erlich *et al.* gave a cost of $3500 per MB—about a quarter of the cost estimated by Goldman *et al.* [19], but they expected to use a more cost-effective approach for DNA synthesis as they developed a powerful error-correcting algorithm that tolerates base errors and losses. Very recently, Lee *et al.* showed a proof-of-principle enzymatic DNA synthesis scheme, which did not achieve single-base precision, but was still sufficient for complete information retrieval and showed a strong cost advantage over traditional phosphoramidite synthesis [37]. In addition, this synthesis scheme also supports a larger storage volume (~500 to several thousand bases per synthesis) at a higher speed. However, in their implementation, the amount of data stored was extremely limited (144 bits) and whether this approach can be scaled up remains to be tested. Advanced coding and decoding algorithms may ultimately lift the technical requirements on synthesis and sequencing and enable production-grade DNA storage. In addition, storage-specific read and write methods may be developed outside the current synthesis and sequencing frameworks. Writing by the massive assemblage of premade oligonucleotides in a way similar to movable-type printing, for example, has recently been claimed to reach a 1 TB/day storage speed.

Random access is another function necessary for information storage purposes. PCR is typically performed using specific primers to obtain selective information stored in DNA. For long-chain DNA storage, PCR with appropriate primers upstream and downstream of the desired information will suffice. However, for oligo DNA storage systems, the entire library needs to be sequenced and assembled before fragmental information can be acquired. Based on powerful error correction codes and algorithmic design, Organick *et al.* developed a framework to minimize the amount of sequencing required to obtain specific data in an oligo library [18]. They managed to retrieve 35 files (with a total size >200 MB) independently without errors. According to their estimates, the method could be extended to an oligo library with a few TBs of storage capacity. It is worth mentioning that the work of Organick *et al*. is also an attempt to store the largest amount of data in DNA molecules so far (at the time of writing in 2019).

Finally, techniques to erase and rewrite information in DNA remain to be developed. Existing DNA storage methods support one-time storage only and thus are suitable for information that does not need to be modified, such as government documents and historical archives. However, the continuous development of synthetic biology has shown the possibility of solving this problem. Artificial gene circuits with stable DNA encoding functions have been designed [70–73,78–81]. For example, using a ‘Set’ system of recombinant enzymes and a ‘Reset’ system of integrase and its excision partner, a controllable and rewritable switch could be implemented [76].

Carbon-based storage

Thanks to the rapid development of DNA manipulation technologies, DNA has become a promising new storage medium. However, other types of polymers may also be used in the field of information storage. Most of them are organic polymers, which, together with DNA molecules, constitute a novel carbon-based storage system different from traditional silicon-based storage.

Like DNA, proteins are an indispensable class of molecules in living systems. Their heterogeneous composition shows potential usage for information storage. However, such attempts are currently focused on the state of the protein rather than its amino acid sequence. For instance, a protein adopting two different states may encode 0 and 1, and information may be stored by switching and stabilizing the states by specific means. A typical example is a photo-switchable fluorescent protein, which changes color when absorbing photons of a particular wavelength [87,88]. Despite its high controllability, the information density is limited to 1 bit per molecule.

In theory, any heterogeneous polymer may serve the purpose of information storage as long as its component monomers can be handled with precision. Current attempts include DNA template guided incorporation of nucleic acid derivatives or small peptides into self-replicating biopolymers [89–91]. In recent years, the discovery of six non-natural nucleic acids that are able to form stable DNA duplex structures and even carry

on genetic information suggests their use for DNA storage [92,93]. In addition to biopolymers, the synthesis of high-molecular-weight polymers such as polyamides and polyurethanes by precise sequence control methods has also been reported in many studies [94–96]. Unfortunately, the read and write techniques for these polymers are far less mature than DNA synthesis and sequencing at the present time. For example, sequencing of synthetic polymers relies on more general analytical methods such as MS/MS and NMR [97–99]. Interestingly, single-molecule nanopore sequencing is expected to be a powerful tool for reading information in synthetic polymers [100,101].

With more types of monomers able to be integrated, synthetic polymers may exhibit higher self-information and thus storage capacity. In addition, it may be more amenable to certain storage functions such as data erasure and rewriting. On a different scale, composite encoding has been applied to information storage. By using mixtures of nucleic acids or metabolites, one can potentially augment coding capacity in the continuous compositional space of components [102,103].

Taken together, synthetic polymers hold great promise for molecular information storage in non-living systems. With the development of sequence control and acquisition technologies, biological and synthetic polymers may form a new framework of carbon-based storage in the future and gradually replace traditional silicon-based storage systems in specialized or general applications.

Acknowledgements

We thank Ming Ni and Yue Shen from BGI-Shenzhen for constructive discussions.

Contributor Information

Yiming Dong, Center for Quantitative Biology and Peking-Tsinghua Center for Life Sciences, Peking University, Beijing 100871, China.

Fajia Sun, Center for Quantitative Biology and Peking-Tsinghua Center for Life Sciences, Peking University, Beijing 100871, China.

Zhi Ping, Academician Workstation of BGI Synthetic Genomics, BGI-Shenzhen, Shenzhen 518083, China.

Qi Ouyang, Center for Quantitative Biology and Peking-Tsinghua Center for Life Sciences, Peking University, Beijing 100871, China; The State Key Laboratory for Artificial Microstructures and Mesoscopic Physics, School of Physics, Peking University, Beijing 100871, China. Long Qian, Center for Quantitative Biology and Peking-Tsinghua Center for Life Sciences, Peking University, Beijing 100871, China.