DNA as data storage

PHG FOUNDATION

making science work for health

Summary

- DNA is being actively researched and developed in the public, academic and private sectors as a storage device for digital (binary) information
- DNA is stable over long time periods and can store information at high density
- Challenges to be resolved include the cost and speed of writing information onto DNA, and information retrieval
- More specific exploration of the use of DNA for health data storage is needed
- While DNA is not yet used routinely for data storage, it has potential to address the challenge of archiving ever increasing volumes of digital data.

DNA as a storage device – also commonly referred to as 'DNA Digital Data Storage' or 'DNA storage' – is an approach for storing digital information in DNA molecules. It is analogous to storing data on a computer hard drive, but instead of using magnetic or electronic technology, data are stored using chemical technology in the form of DNA molecules.

In organisms a sequence of the four DNA nucleotides – A, T, G, C – code for a set of instructions to make a protein. These four nucleotides, or letters, can also function as a coding system analogous to the binary digital code used by computers to represent information such as letters, digits, or other characters. In most cases, DNA has been used to represent binary code itself. The first proof-of-concept example was in 1988 and since then, the size and complexity of data stored in DNA has increased, including text, images, audio, and a computer operating system.

The data challenge

Global data production is expected to soon exceed the capacity of currently available storage methods and DNA digital data storage is seen as a possible solution to meet these demands. Although this is a technology based on molecular science, it has potential application in any field where data storage and archiving are needed. Current work in this area is not yet looking specifically into healthcare applications, but on developing the storage technology.

WYNG FOUNDATION

The views set out in this briefing note do not necessarily represent those of the WYNG Foundation

DNA storage approaches are being researched and developed in both the public and private sector, for example:

- The University of Washington's Molecular Information Systems Lab, in partnership with Microsoft Research, is exploring methods of access of information from synthetic DNA
- Scientists at Peking University's Centre for Quantitative Biology have been developing a method of storing data in the genome of an extremophile bacterium
- Companies such as Twist Bioscience are working on methods for synthesising DNA

Storing and retrieving data in DNA

Storing and retrieving data from DNA generally involves the following steps:

1. **Encoding.** Information is encoded by assigning unique combinations of DNA nucleotides to specific binary bits. To prevent repetitive sequences being encoded, which can be hard to sequence accurately, it is common for two or more nucleotides to be used to code for one bit of information.
2. **DNA synthesis (writing data).** DNA is constructed using A, T, G and C nucleotides in a sequence corresponding to the encoded data.
3. **DNA storage.** DNA is protected from degradation and errors in its code – typically caused by environmental factors like UV, moisture, and oxygen – by encapsulation and storage at stable temperatures, ranging from room temperature to -80°C.
4. **Retrieval and DNA sequencing (retrieving and reading data).** DNA is usually amplified through polymerase chain reaction (PCR) and then sequenced to determine the order of nucleotides. New methods, such as 'random access' retrieval, are being developed to avoid the costs of sequencing the whole pool of DNA when only a subset of information is required. Approaches can use 'primers' to selectively target and amplify specific DNA data sequences, or 'barcode' sequences that tag specific sections of the DNA for more rapid retrieval.
5. **Decoding.** Sequenced DNA is converted or 'decoded' into the binary code representing the original data. Before decoding, error correction algorithms can be used to identify and correct errors that might have been introduced in the DNA during the synthesis, preservation or sequencing steps.

Variations in these steps include:

- **DNA microarray / microchip technology** has been used perform DNA synthesis directly onto chips which can reduce the time taken to write data onto DNA
- **CRISPR-Cas** gene editing, which uses a modified CRISPR system to encode data directly into bacterial (*E. coli*) cells, bypassing the need to synthesise DNA
- **'DNA of things'**, which integrates synthetic DNA into everyday objects, meaning they contain information about them, including how to produce them. In healthcare, a 'DNA of things' approach to storing data within personalised medical devices or implants, as a back-up to medical records, could be explored

Other potential uses in healthcare include handling large data volumes, such as medical records and research data, and securing backups for critical medical records vulnerable to cyber-attacks.

What are the enablers and barriers?

While there are many enablers to using DNA data storage, there are also challenges to its use beyond research, including in healthcare:

The enablers are:

- **Global data storage demand:** There is a growing need for high-fidelity, durable, compact data storage due to increasing global data volumes, estimated around 33 zettabytes (33 trillion gigabytes) by 2025 [1]
- **Long-term stability:** Synthetic DNA's potential to remain stable for thousands of years lends itself to historical record archiving, in contrast to current storage technologies such as magnetic tape where data need to be transferred if long term (>50 years) preservation is required
- **High density storage:** DNA can store vast amounts of information in a very small volume, orders of magnitude more than current storage media which are reaching their density limit
- **Ease of storage:** DNA's stability and compactness reduce physical storage overheads compared to conventional systems
- **Investment:** Strong interest and funding from governments, intelligence agencies, and private or public research institutions

The barriers are:

- **Time constraints:** Slow and costly DNA storage and retrieval processes, especially DNA synthesis and sequencing, make the everyday use of DNA data storage currently impractical
- **Error rates and data integrity:** Potential for errors during the different stages of the data storage process, such as DNA synthesis and sequencing, can affect data integrity, the impact of which could be minor to severe, depending on the intended use
- **Cost of DNA synthesis and sequencing:** for writing and reading data from DNA, makes these steps far more costly compared to conventional data storage approaches. Estimates for storing a gigabyte (GB) of data using DNA as storage medium are ~$130/GB compared to $0.015 - $0.02/GB for cloud storage [2,3]
- **Specialist skills:** such as expertise in data science, bioinformatics, and laboratory protocols are required for storing data on DNA
- **Lack of automated, integrated system:** Unlike current digital storage systems (e.g. computers), the multiple steps involved in storing data on DNA are not integrated since the technologies underpinning DNA data storage were not originally designed for this purpose

The utility of DNA data storage for any type of data requiring regular or rapid access is currently limited by the slow speeds of DNA synthesis and sequencing steps relative to electronic information processing. However, the very long stability of DNA suggests its use for long term archival storage, where data are not frequently accessed, is much more feasible.

Considerations

Legal and regulatory. If DNA is used to store information about living individuals it is feasible that the stored DNA itself will be treated as 'personal data'. However, this will depend on whether the DNA material is treated as 'information' in its own right or as a 'source of information'. It will also turn on how likely it is that someone may be to retrieve, sequence and identify an individual from that information.

Social and ethical. Questions still to be explored more thoroughly are around access to DNA synthesis technology and biosecurity and biosafety concerns should DNA data storage become widespread. While ethical considerations will depend on implementation, possible concerns could include:

- **(In)Equity in historical records:** Uneven access to DNA data storage might lead to biased records and power imbalances in history preservation.
- **Data security:** Threats to DNA-stored data and potential hacking are unexplored. This includes the 'DNA of things' which could theoretically enable covert information transfer.

Evidence gaps

- **Implementation model:** There is a lack of a practical roadmap to transition DNA data storage from proof-of-concept to a functional, accessible service. Details about service provision, user access, interfaces, and data submission/access are unclear. Industry standards, a shared ecosystem for technology adoption and user-friendly tools are needed.
- **Healthcare feasibility:** If a practical archival model existed, evidence gaps would shift to assessing DNA data storage's viability for healthcare. This includes feasibility for medical data, error tolerance, and its applicability for healthcare applications.

Looking ahead

Internationally there is great interest in this technology, with the DNA Data Storage Alliance exploring implementation issues, investment from the private sector, and funding initiatives in the EU and US. It is likely that DNA data storage will be in regular use within the next 20 years, conceivably much sooner. While it has potential for revolutionising data storage with its vast capacity and longevity, substantial challenges such as high costs and slow data read/write speeds (relative to conventional methods) must be overcome. Using DNA for storing health data needs its own specific investigation.

References

1. Ionkov, L., and Settlemyer, B. (2021) DNA: The Ultimate Data-Storage Solution. Scientific American https://www.scientificamerican.com/article/dna-the-ultimate-data-storage-solution/
2. DNA Storage Alliance (2021) Preserving our digital legacy: an introduction to DNA storage. https://dnastoragealliance.org/dev/wp-content/uploads/2021/06/DNA-Data-Storage-Alliance-An-Introduction-to-DNA-Data-Storage.pdf
3. Hale, C. (2023) How much does cloud storage cost? Tech Radar. https://www.techradar.com/features/how-much-does-cloud-storage-cost

Author: Dr Sobia Raza

Published: February 2024

UNIVERSITY OF CAMBRIDGE