# From megabytes to molecules: The usage of DNA computing for storage and its implications on the future

By Urvi Shetty

## Author Biography

Urvi Shetty is currently a sophomore studying in Karnataka, India. She has always had a passion for computer science, neuroscience and mathematics alongside the musical instruments she plays including violin, keyboard, guitar, and many more. As someone with a wide range of interests, she enjoys exploring intersectionalities between fields, always seeking new topics to research. Due to this curiosity, she takes courses on a variety of topics including coding, robotics, research, maths and more, always searching to build on her current knowledge.

## Abstract

It is more important now than ever to develop more effective data storage techniques in the constantly advancing field of computers. Due to their limits in terms of durability, information density, and physical space requirements, current storage technologies, such as magnetic and optical media, are finding it difficult to keep up with the global demand for data storage, which is expected to exceed 175 trillion gigabytes by 2025.
Nature's own method of storing genetic information for a long time, deoxyribonucleic acid (DNA), provides great data density, stability, and endurance, making it a viable solution to address this issue.

Through a comprehensive analysis of existing literature, this paper aims to provide a nuanced understanding of DNA computing's current state - mainly its trajectory towards becoming a viable alternative to traditional storage methods - highlighting the history, current developments/challenges, and future prospects.

*Keywords: DNA, DNA Storage, DNA synthesis, DNA sequencing, nanopore sequencing, storage, encoding algorithms, chemical synthesis, enzymatic synthesis, binary, nucleotides*

## Introduction

In the rapidly evolving landscape of computing, the invention of more efficient data storage methods has become increasingly critical. Data storage can be described as the technological means by which digital data is retained and archived for use on a computer or other related device (1). With the huge increase in global data, the demand for data storage worldwide is predicted to reach $1.75 \times 10^{14}$ GB (175 trillion) by 2025 (2) which our current storage methods struggle to keep pace with. This also means even more physical space will be needed for memory.(3) A vast majority of our data is currently stored on magnetic media such as hard drives and tapes, with some data also stored on optical media like CDs and DVDs. These storage methods present us with several disadvantages including low durability (<50 years), low information density, and the need for specialised readers (2). Given these limitations, researchers are exploring various alternatives to meet the constantly rising data storage demands, and as nature's own system for reliable, long-term storage of genetic information - deoxyribonucleic acid (DNA) offers remarkable data density, stability, and longevity, making it a promising potential solution. No organisms or cells are used for DNA data storage; rather, the DNA is synthesised through controlled chemical processes (4) which will be discussed later in this paper.

## Process

DNA is made up of two chains of smaller molecules called nucleotides, coiled around each other to form a double helix. Each chain contains a string of nucleotide monomers, also referred to as bases: adenine (A), thymine (T), cytosine (C) or guanine (G)(5). The two chains are held together by bonds formed between the bases of the opposite chain. These bases are complementary to a specific type of base, which means A only forms bonds with T, and C only forms bonds with G. The process of using DNA as a storage method undergoes many stages, which can be broken down into encoding, synthesis, storage, retrieval, sequencing and decoding (4, 5).

## Encoding

Computers represent all the data they work with in the form of 0s and 1s (binary). The process of converting these 0s and 1s into sequences of bases (A,T,C,G) is called encoding. There are various methods of encoding that have been developed.

### Modern day encoding algorithms

The encoding method we use directly affects (and is affected by) the overall quality of the data storage system including factors such as data density, cost, and data reliability (5). Two encoding algorithms that are regularly used include either assigning two binary digits to each base, or creating a library of small oligonucleotides (short molecules of DNA) representing letters, symbols, or numbers where the original 0's and 1's are encoded to the letters in the alphabet.

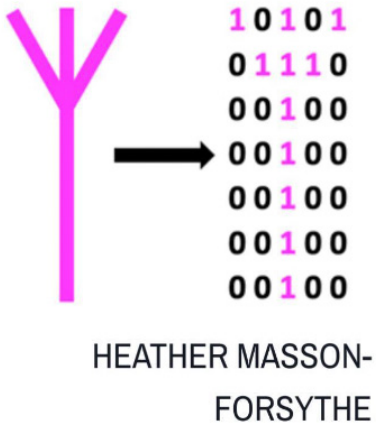
In the first method, each base would be assigned 2 binary digits, for example:

A = 00
C = 01
G =10
T = 11

Thus, a number such as 0011011001 could be encoded as ATCGC, which would then be synthesised (discussed further in section 4) (5).

## Microvenus

Figure 1
The microvenus binary code




Note: Source of picture: Heather Masson-Forsythe, https://www.asbmb.org/asbmb-today/science/101322/artists-technologists-bring-data-storage-to-life.

The first attempt at using DNA for storage purposes was for a collaboration of art and science, Microvenus (1996) (6). Their idea was to incorporate data in DNA, more specifically a living bacteria (e. Coli). The information they encoded was the ancient Germanic rune for “life” and “female Earth”, within a 5x7 or 7x5 figure (35 bits of data). The encoding strategy they used was assigning phase-change values to each of the bases, where each base indicates how many times each binary digit is to be repeated. This can be written as:

C = X, T = XX, A = XXX, G = XXXX
where X is a binary digit.

Thus, the first five digits of the microvenus binary code—101010—would be written as CCCCC (as the binary digits occur only once before changing), while the next few digits—0111000—would be CAA. By using this encoding method, the thirty-five bit binary map would translate to only eighteen bases: CCCCCCAACGCGCGCGCT. An additional sequence CTTAAAGGGG was added to the beginning of the DNA sequence as a clue into their assigned values, thus giving us a sequence with a final length of 28 nucleotides (which were then incorporated into the living cell using recombination molecular biology technique) (7).

# Synthesis

DNA synthesis is the creation - naturally or chemically - of DNA molecules. In nature, they are synthesised through the process of DNA replication, often as a part of cell division. In order to synthesise DNA in an efficient and cost-effective manner, the methods we devise are constantly undergoing improvements. This is also due to the vast amount of applications of DNA synthesis, including genetic modification, protein engineering, molecular tools and much more (8). Chemical methods, followed by enzymatic and hybrid methods were developed for the step-wise synthesis of DNA (base-by-base or nucleotide-by-nucleotide), which allowed us to generate DNA of longer length and higher complexity than oligonucleotides (9).

## Chemical and Enzymatic synthesis

Chemical synthesis (phosphoramidite synthesis) as well as enzymatic synthesis are both base-by-base synthesis methods. This is done by beginning the process with a base bound by a chemical element called a blocker. The strand is then “de-blocked” and a new base with a blocker is added to the strand. This helps prevent the synthesis of undesired molecules and allows the DNA chain to synthesise in the right order. This process is then repeated to continue synthesising the DNA, while maintaining acceptable error rates (5).

In phosphoramidite synthesis (phosphoramidites are building blocks used in chemical synthesis of oligonucleotides(10)), the oligonucleotides are synthesised from building blocks that replicate natural bases (5). On the other

hand, enzymatic DNA synthesis (EDS) is another method developed as an alternative to phosphoramidite synthesis where instead of solvents and phosphoramidites, they make use of aqueous reagents and enzymes (11). While phosphoramidite synthesis currently dominates the market compared to EDS due to advantages in quality and robustness, EDS is progressing rapidly. DNA printers have already been developed, such as DNA script (the world's first DNA printer powered by EDS), making this technology more accessible and slowly building on its efficiency as a whole by enabling the parallel synthesis of multiple sequences.

The ideal path to take in the future might be a hybrid format, which will be difficult but essential to work towards in order to build up to more efficient, affordable, and accurate DNA synthesis methods.

## Storage/Preservation

The storage environment of DNA has a significant impact on its stability. Environmental factors play a huge role in the preservation of DNA, as DNA is known to be degraded by interactions with water, UV irradiation, pollutants, etc. The environmental factors that need to be controlled can be listed by: temperature, water, UV irradiation, oxidation, and pH value.

| *Environmental Factor:* | *Relation to DNA:* |
|---|---|
| Temperature | DNA is stored in many ways: cold storage at –4 °C, freezing DNA at –20 or –80 °C, cryopreservation in liquid nitrogen, etc, the most common approach being to freeze the DNA at ultralow temperatures.<br><br>However, studies found that cryopreservation of DNA could cause damage.A study revealed that DNA in a buffered solution kept at –20 or –80 °C gradually deteriorated, but DNA kept at 4 °C remained stable. Additional research also found that the slower freezing of DNA was more damaging than rapid freezing. |
| Water | We should minimise its exposure to moisture. Water is a huge degradation factor - the half-life of DNA exposed to moisture degrades significantly. |
| UV irradiation | UV light can induce DNA strand breaks, cross-linking, nucleobases modification, and more. However, newer discoveries (from reference 16) lead us to believe that DNA may be specially designed to protect itself from UV light in the future. |
| Oxidation | Reactive oxygen species (often referred to as ROS; highly reactive chemicals formed from oxygen, water, and hydrogen peroxide), which are another main factor of DNA degradation during storage |
| pH value | Neutral or slightly alkaline pH is suitable to preserve DNA. |

Note: Source of data: Tan, X., Ge, L., Zhang, T., & Lu, Z. (2021). Preservation of DNA for data storage. Russian Chemical Reviews, 90(2), 280.

## Retrieval

Retrieval is the extraction of DNA molecules from a DNA archive to prepare it for sequencing. During the encoding process, each DNA can be designed to have its own unique 'barcode' block to identify it. By doing this, a specific sequence can be read when required as in modern random-access memory (RAM). Then, when the data needs to be decoded, the DNA can be selectively retrieved and amplified (using polymerase chain reactions [PCR] - a technique to rapidly produce a large number of copies of a specific segment of DNA) from the pool, using specific primers (a short, single-stranded nucleic acid) corresponding to the barcode block (12).

## Sequencing

DNA sequencing is the term for the process of detecting the order of bases in a strand of DNA. The two main methods used for sequencing data are sequencing-by-synthesis (SBS) and nanopore sequencing (4). Both of these fall under the category of next generation sequencing (NGS).

### Sequencing-by-synthesis

SBS determines the sequence of bases in a DNA sample using an engineered DNA polymerase (the enzyme that catalyses the synthesis of DNA) to synthesise a copy of a single strand of DNA (13). This step is generally referred to as "library preparation". It then indirectly identifies the sequence of the bases by reconstructing the strand back into its (double-stranded) DNA form using a polymerase to incorporate the new strand of complementary bases to the strand. Every "incorporation" generates a signal that is then detected and used to determine the sequence of bases (4).

### Nanopore Sequencing

Nanopore sequencing detects the sequence of bases directly by passing a strand of DNA through a pore in a kind of membrane surrounded by an electrolyte solution. By measuring the disturbances in the ionic current as the strand passes through the pore (only a few nm wide), we can determine the sequence of bases quite efficiently (4, 5).

There is a need for efficient and cheap sequencing. One nanopore working at 10 ms/base would take about 20 years to sequence a human genome. Ten thousand pores would take less than a day for a genome, and a million pores would only take 10 minutes. This highlights not only the potential of nanopore sequencing, but also the important of parallel sequencing (14)

### Decoding

The final step in DNA storage is decoding, where the stored DNA sequences are read and translated back into digital data. Once the sequence of bases is obtained, bioinformatics tools are used to map these sequences back to their corresponding binary values - quite literally reversing the encoding process. Error correction algorithms are often applied here to ensure data integrity, and watch out for sequencing errors. Through this step, we finally achieve a digital reconstruction of the original data successfully retrieved from its DNA form.

## Conclusion

The ever-increasing global data demands leads us to keep working on innovative data storage solutions. Traditional storage methods come with many limitations such as durability, information density and more. DNA storage can serve as a great alternative, providing us with unparalleled data density, stability, and longevity.

However, the current cost of DNA synthesis and sequencing is quite high, making DNA storage quite expensive compared to traditional methods like magnetic and optical storage. The process of encoding, synthesising, retrieving, sequencing, and decoding DNA is also relatively slow and although advancements are being made, DNA synthesis and sequencing can still produce a lot of errors, which require us to develop sophisticated error correction algorithms.

As we move forward, and work towards addressing the challenges of data integrity, cost and scalability, DNA data storage still stands as a viable solution for the global data storage issue. It gives us a promising future where our data storage solutions are as resilient and enduring as the genetic materials that make up our own bodies.

## Acknowledgment

I would like to express my sincere gratitude to the mentors from the Ivy Scholars Research Bootcamp for their invaluable guidance and support. I am also deeply appreciative of my family- especially my parents, for their unwavering support for all my work.

## References

(1) Thomas Alsop, Jan 15, 2024, Data Storage - statistics & facts, Statista

(2) Doricchi, A., Platnich, C. M., Gimpel, A., Horn, F., Earle, M., Lanzavecchia, G., Cortajarena, A. L., Liz-Marzán, L. M., Liu, N., Heckel, R., Grass, R. N., Krahne, R., Keyser, U. F., & Garoli, D. (2022). Emerging Approaches to DNA Data Storage: Challenges and Prospects. ACS nano, 16(11), 17552–17571. https://doi.org/10.1021/acsnano.2c06748

(3) Testing memory downsizing limits. (2019). Nature Nanotechnology, 14(1), 1. https://doi.org/10.1038/s41565-018-0355-0

(4) The DNA Data Storage Model. (2023, July 1). IEEE Journals & Magazine | IEEE Xplore. https://ieeexplore.ieee.org/document/10154188

(5) DNA Data Storage Alliance, Preserving Our Digital Legacy: An Introduction to DNA Data Storage, June 1, 2021

(6) Davis, J. (1996). Microvenus. Art Journal, 55(1), 70–74. https://doi.org/10.2307/777811

(7) Thomas Heinis, Roman Sokolovskii, and Jamie J. Alnasir. 2023. Survey of Information Encoding Techniques for DNA. ACM Comput. Surv. 56, 4, Article 107 (April 2024), 30 pages. https://doi.org/10.1145/3626233

(8) Kosuri, S., & Church, G. M. (2014). Large-scale de novo DNA synthesis: technologies and applications. Nature methods, 11(5), 499-507.

(9) Hoose, A., Vellacott, R., Storch, M. et al. DNA synthesis technologies to close the gene writing gap. Nat Rev Chem 7, 144–161 (2023). https://doi.org/10.1038/s41570-022-00456-9

(10)Phosphoramidites for Nucleic Acid Synthesis | Thermo Fisher Scientific - IE. (n.d.). https://www.thermofisher.com/in/en/home/life-science/oligonucleotides-primers-probes-genes/phosphoramidites.

(11) Verardo, D., Adelizzi, B., Rodriguez-Pinzon, D. A., Moghaddam, N., Thomée, E., Loman, T., Godron, X., & Horgan, A. (2023). Multiplex enzymatic synthesis of DNA with single-base resolution. Science Advances, 9(27). https://doi.org/10.1126/sciadv.adi0263

(12) Eojin Yoo, Donghui Choe, Jongoh Shin, Suhyung Cho, Byung-Kwan Cho, Mini review: Enzyme-based DNA synthesis and selective retrieval for data storage, 2021, ISSN 2001-0370, https://doi.org/10.1016/j.csbj.2021.04.057.

(13) Mathias Uhlen, Stephen R. Quake, July 22 2023, doi: 10.1016/j.tibtech.2023.06.007•

(14) Bayley H. (2015). Nanopore sequencing: from imagination to reality. Clinical chemistry, 61(1), 25–31. https://doi.org/10.1373/clinchem.2014.223016

(15) Masson–Forsythe, H. (2022, October 13). Artists, technologists bring data storage to life. https://www.asbmb.org/asbmb-today/science/101322/artists-technologists-bring-data-storage-to-life

(16) Akash, A., Bencurova, E., & Dandekar, T. (2024). How to make DNA data storage more applicable. Trends in biotechnology, 42(1), 17–30. https://doi.org/10.1016/j.tibtech.2023.07.006

(17) NCI Dictionary of Genetics Terms. (n.d.). Cancer.gov. https://www.cancer.gov/publications/dictionaries/genetics-dictionary/

(18) Clelland, C. T., Risca, V., & Bancroft, C. (1999). Hiding messages in DNA microdots. Nature, 399(6736), 533–534. https://doi.org/10.1038/21092

(19) Xin Tan et al 2021 Russ. Chem. Rev. 90 280, DOI 10.1070/RCR4994

(20) Barlev, A., & Sen, D. (2018). DNA's encounter with ultraviolet light: an instinct for self-preservation?. Accounts of Chemical Research, 51(2), 526-533

(21) Panda D, Molla KA, Baig MJ, et al. DNA as a digital information storage device: hope or hype? 3 Biotech. 2018 May;8(5):239. DOI: 10.1007/s13205-018-1246-7. PMID: 29744271; PMCID: PMC5935598.