



STORING BIG DATA IN TINY DNA MOLECULES

Daniella Bar-Lev^{1*}, Inbal Preuss^{2,3†}, Leon Anavy^{2,3} and Eitan Yaakobi¹

¹The Henry and Marilyn Taub Faculty of Computer Science, Technion, Haifa, Israel

²Arazi School of Computer Science, Reichman University, Herzliya, Israel

³Department of Computer Science, Technion – Israel Institute of Technology, Haifa, Israel

YOUNG REVIEWERS:



**EISNER
FAMILY**

AGES: 11–15

ZALMAN

ARAN

JUNIOR



**HIGH,
HOLON:
8TH GRADE**

AGES: 13–14

How many pictures did you take this week? Did you watch videos on YouTube, or maybe a series on Netflix? All of these activities rely on digital information stored on your smartphone, computer, or in the cloud, ready for you to access whenever you want. As people create more and more data, the technology we use to store it has to change to handle new and challenging problems. One promising solution is the use of synthetic DNA molecules. DNA is a natural material found in most cells of all living organisms, and it stores the organism's genetic information—huge amounts of biological data. Today, scientists know how to make artificial DNA molecules in the lab. This means it is possible to use the enormous storage potential of DNA to store digital information instead of just biological data. This article explains how scientists are combining biology and computer science to create artificial DNA molecules that can store

large amounts of digital information for a long time, in a way that is good for the environment.

THE DATA-STORAGE CHALLENGE

The amount of digital information being created today is growing at a dizzying pace. According to some estimates, by 2025, there will be **five times more digital data** than there was in 2018 (Figure 1), and it all needs to be stored somewhere. This huge rise in data is driven by several things, including the use of smartphones, social media, and new technologies like **artificial intelligence (AI)**. These technologies create various types of data, such as personal information, social media posts, business records, medical statistics, sensor data, and other types of digital information stored on computers.

ARTIFICIAL INTELLIGENCE (AI)

Computer systems or machines that mimic human intelligence to perform tasks.

Figure 1

The expected amount of digital information worldwide, from 2010 to 2025. Information is measured in Zettabytes, which are a unit of digital information equal to a billion terabytes (1,000,000,000 TB), used to measure massive amounts of data. One terabyte (TB) is about the storage space on a typical laptop, so one zettabyte could hold data from a billion laptops.

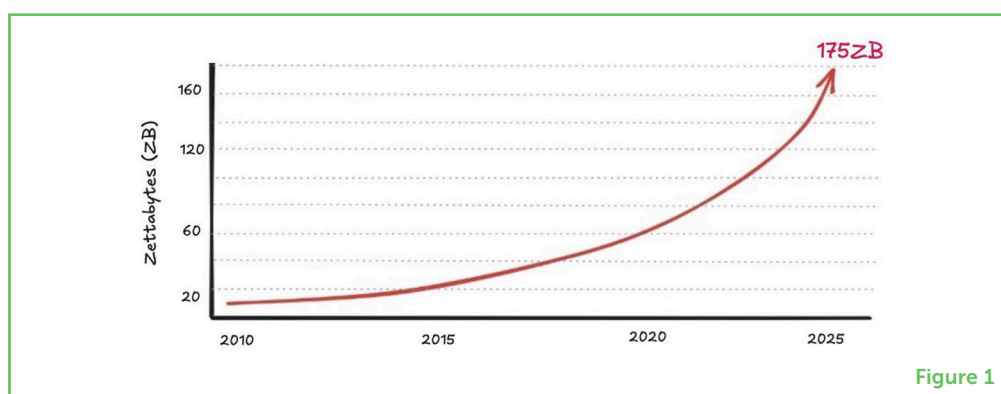


Figure 1

As technology becomes more advanced, we create more and more information that needs to be stored and accessed easily. It might seem like having more information is a good thing, but there is a problem. The challenge is figuring out how to store this growing stockpile of information. Specifically, we are approaching a point where our current storage technologies cannot keep up with the amount of data that is being created. This has far-reaching implications because our technological world depends on the ability to store digital information effectively.

You might have heard people say that saving files on the cloud gives you endless storage, as if the cloud were an infinite, invisible space. But this is not true. In reality, the cloud is not invisible; it is made up of huge storage centers, sometimes as big as several football fields. These centers hold thousands, or even millions, of computers that store and manage enormous amounts of digital information. Not only do they take up a lot of space, but they also use as much electricity as a mid-sized city. This means that storing information this way is not **sustainable**, so we must find better ways to store data (Figure 2).

SUSTAINABLE

A way of using resources without damaging the environment or using up resources needed by future generations. Sustainable solutions are important for protecting our planet over time.

Figure 2

Meta (formerly Facebook)'s digital information storage center (figure credit: Meta).



Figure 2

SYNTHETIC DNA

DNA molecules created in the lab, which are not derived from natural sources.

DNA SEQUENCING

A laboratory technique used to determine the exact order of the four letters that make up a DNA molecule. DNA sequencing allows scientists to “read” the information stored in DNA.

DNA-BASED DATA STORAGE

When we talk about storing information, there are a few important things to think about. First, the storage solution should be able to hold a lot of information. Second, it should be good for the environment. And third, it should last a long time. One exciting idea that can do all of these is using **synthetic DNA** molecules to store digital information. But how does this work? And what makes DNA a good choice for storage?

DNA is a special molecule found in the cells of every living thing, from tiny bacteria to humans. DNA holds all the genetic information needed to create life. You can think of a DNA molecule as a long chain made up of four “building blocks”: adenine (A), guanine (G), thymine (T), and cytosine (C). The order in which these four building blocks are arranged in a chain is unique to each living organism. For a long time, scientists have known how to “read” the exact order of the four building blocks in a DNA molecule, using a process called **DNA sequencing** [1]. More recently, scientists have learned how to design and create synthetic DNA in the lab by putting the building blocks together in almost any order they like.

WHY DNA?

Compared to other ways we store digital information today, DNA molecules have some important advantages. First, although DNA molecules are tiny, they can hold an enormous amount of information. Using DNA for information storage would mean we would not need the huge storage centers used today, which take up a lot of space and energy. For example, if we store many DNA molecules containing digital data in a storage container the size of an average person, this container could hold 150 billion terabytes of data! That is an incredible amount of storage, equivalent to about 150,000 storage centers.

BINARY LANGUAGE

The language used for computers, in which information is represented by combinations of only two numbers, 0 and 1.

The second advantage is that DNA has a stable structure that can last a very long time. In fact, scientists can even sequence DNA from fossils of creatures that roamed the Earth hundreds of thousands of years ago. For this reason, digital information could be stored and preserved in DNA for a very long time, unlike our current storage technologies, which usually last only 3–30 years. Finally, DNA does not need a constant source of electricity to keep the data stored. This could save a lot of energy, making DNA-based data storage an environmentally friendly solution.

TRANSLATING DIGITAL INFORMATION INTO DNA LANGUAGE

How do we take digital information, like a music file, video, or image, and put it on a tiny DNA molecule? The secret lies in the language! Digital information is stored on computers using the **binary language**, which consists of only two digits: 0 and 1. On the other hand, DNA's genetic code uses a language made up of four letters: A, C, G, and T. If we can translate between these two languages, we could store a lot of digital information in tiny DNA molecules. Essentially, we translate information currently in binary language into DNA language and then create the matching synthetic DNA molecules and store them.

To translate one language into another, we need common ground to connect them. If we look at the binary language, we can take the numbers 0 and 1 and pair them together into four combinations:

00

01

10

11

These four combinations work perfectly because they can match the four letters in the DNA language. So, we can replace each binary pair with a DNA letter:

00 becomes **A**.

01 becomes **C**.

10 becomes **G**.

11 becomes **T**.

For example, let us take the binary sequence 00010011111001001. We can split it into pairs and mark each pair with a color like this:

00 01 00 11 11 10 01 00 11

Now, we translate each binary pair into a DNA letter:

A C A T T G C A T.

Remember, this is just a simple example. In real life, scientists use more advanced methods to ensure that the information stays safe and accurate, even if there is a mistake in the system [2].

STORING DIGITAL INFORMATION IN DNA

After the binary code is translated into DNA language, there are several more steps in the data-storing process [3], which are summarized in Figure 3.

Figure 3

Using synthetic DNA for information storage. Step 1: The binary language that makes up the digital files is read. Step 2: The digital information, written in binary language, is translated into DNA language. Step 3: The information is stored as synthetic DNA by creating the desired DNA molecules in the lab. Step 4: Later, the information can be accessed by reading the stored DNA molecules, which gives us words in DNA language. Step 5: Decoding and translating the words from DNA language back into the original digital file, which is done with the help of a computer program (figure created with BioRender.com).

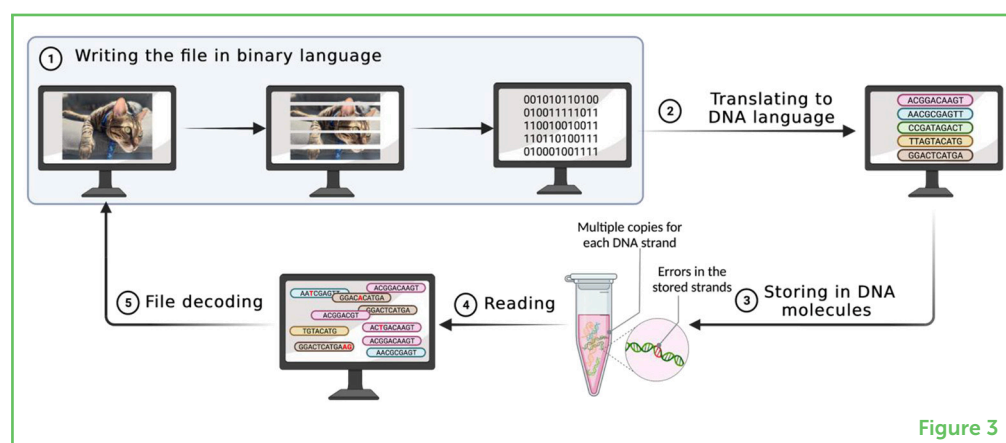


Figure 3

First, due to technology limitations, the DNA molecules we can make in the lab are short, which is a problem, because our files can be much larger than what can be stored in a single molecule. Therefore, we must cut our digital files into smaller sections before each piece is written in binary language. The binary information is then translated into DNA language as we just explained, with the help of a computer program. Finally, in the lab, we create the desired DNA molecules using advanced equipment. Interestingly, when we make one molecule, a lot of nearly identical copies are produced at the same time. All of these molecules are stored together in storage containers without any specific order. That is, DNA molecules are stored in a pool in a way that is similar to shredding multiple copies of the same book into a million fragments and mixing all the book fragments in a box. Sometimes, errors can happen during this process, which can affect some of the molecules. This can be thought of as losing some of the book fragments, and having typos in some of the remaining fragments (e.g., "Cat" becomes "Cut").

RETRIEVING DATA FROM DNA STORAGE

At this point, the digital information has been successfully transferred into synthetic DNA molecules, which can be stored for a very long time. At any time in the future, we can retrieve the data from these molecules and transform it back into the original digital files. The first step in this process is to read the stored DNA molecules using DNA sequencing. This process gives us a collection of words in the DNA language created by DNA's four building blocks. Then, we translate the DNA words back into binary language, which recreates the original digital file. A computer program helps with this step, too.

SUMMARY AND MAIN POINTS TO REMEMBER

In this article, we explored the challenges of storing digital information and introduced an exciting new solution: synthetic DNA molecules. While we still have a long way to go before we can upload information into a DNA-based cloud, there are ongoing efforts to reduce costs and improve the technology. The next time you take a picture or watch a video, consider where that information is stored. Soon, you might be able to save your own digital data on a synthetic DNA molecule, which can hold a huge amount of information for a very long time while also being environmentally friendly. This incredible achievement might be made possible by bridging the worlds of biology and computer science.

ACKNOWLEDGMENTS

This work was supported by the European Union (DiDAX) under Grant 101115134.

AUTHOR DISCLAIMER

Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

REFERENCES

1. Sanger, F., Nicklen, S., Coulson, A. R. 1977. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci.* 74:5463–5467. doi: 10.1073/pnas.74.12.5463
2. Cosman, P. 2015. *The Secret Code Menace*. Winchester: Ransom.
3. Grass, R. N., Heckel, R., Puddu, M., Paunescu, D., and Stark, W. J. 2015. Robust chemical preservation of digital information on DNA in silica with

error-correcting codes. *Angewandte Chemie Int.* 54:2552–2555. doi: 10.1002/anie.201411378

SUBMITTED: 18 February 2025; **ACCEPTED:** 09 May 2025;
PUBLISHED ONLINE: 05 June 2025.

EDITOR: Idan Segev, Hebrew University of Jerusalem, Israel

SCIENCE MENTORS: Janet Benyamin and Yonatan Aizner

CITATION: Bar-Lev D, Preuss I, Anavy L and Yaakobi E (2025) Storing Big Data in Tiny Dna Molecules. *Front. Young Minds* 13:1579129. doi: 10.3389/frym.2025.1579129

CONFLICT OF INTEREST: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

COPYRIGHT © 2025 Bar-Lev, Preuss, Anavy and Yaakobi. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](#). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

YOUNG REVIEWERS

EISNER FAMILY, AGES: 11–15

We, the children of the Eisner family love science and technology. Shir, the eldest, 15 years old, loves to read. Uri, the middle, 13 years old, loves to learn Torah and is interested in science and fantasy. Yishai, the youngest, 11 years old, loves physics and Dungeons & Dragons.

ZALMAN ARAN JUNIOR HIGH, HOLON: 8TH GRADE, AGES: 13–14

We are curious 8th graders from Zalman Aran Middle School in Holon. This year, our entire class took part in the Frontiers for Young Minds program, in which we were challenged with interesting topics, learned how to read a scientific article and reviewed papers. Our activities with Frontiers Science for Young Minds are part of our curriculum, and appear on our end-of-year report card.

AUTHORS

DANIELLA BAR-LEV

Ph.D. student at the Faculty of Computer Science at the Technion–Israel Institute of Technology. Daniella holds a bachelor's degree in computer science, a bachelor's degree in mathematics, and a master's degree in computer science, all from the Technion. Her research focuses on coding solutions and algorithms for DNA-based



storage systems. Daniella believes that collaboration between researchers in different fields is essential to driving innovative research and making a real contribution. Therefore, in her research, Daniella strives to collaborate and share knowledge with researchers in a variety of fields, such as biologists, chemists, AI researchers, and more. In her spare time, Daniella participates in a variety of activities to expose youth, and girls in particular, to the fields of science and engineering. *dbarlev@ucsd.edu



INBAL PREUSS

Ph.D. student at the Faculty of Computer Science at Reichman University and a research assistant at the Technion–Israel Institute of Technology. Inbal holds a bachelor's degree in computer science, and she is currently involved in several research projects in the fields of computational biology, information sciences, and synthetic biology. Her research focuses on developing DNA-based data storage solutions, in collaboration with Reichman University and the Technion. Inbal worked as a programmer for several years at the Microsoft R&D center in Israel, where she was involved in the development of the touch screen and pen for Surface computers. As the first female doctoral student in the Faculty of Computer Science at Reichman University, Inbal also actively promotes the inclusion of women in technology, out of her belief in the power of women and diversified teams to enhance creative vision, leading to technological breakthroughs. *inbalpreuss@gmail.com



LEON ANAVY

Lecturer at Reichman University's School of Computer Science and visiting scholar at Alphabet's Verily Life Sciences, LLC. Leon researches and teaches topics at the intersection of AI, machine learning, and computational and synthetic biology. His research focuses on the development of coding methods for DNA-based data storage and the use of synthetic biology and machine learning to analyze and characterize complex molecular biological processes. After receiving his bachelor's degree in industrial engineering and management and a master's degree in computational biology, Leon completed his doctorate in computer science, all at the Technion. In addition to his academic work, Leon is involved in various initiatives in the fields of education aimed at integrating disadvantaged populations into the sciences and high-tech fields.



EITAN YAAKOBI

Associate professor at the Faculty of Computer Science at the Technion–Israel Institute of Technology. He holds a bachelor's degree in computer science and a bachelor's degree in mathematics from the Technion, a master's degree in computer science from the Technion, a Ph.D. in electrical engineering from the University of California, San Diego, and a postdoctoral fellowship from Caltech and the University of California, San Diego. Eitan has won a variety of awards for his research, including the Young Researcher Award from the Marconi Society. His interests include coding theory and information theory with applications for information storage components, DNA-based information storage, information storage and retrieval, and private information retrieval. In his spare time, Eitan enjoys spending time with his family, doing yoga, and surfing.

[†]These authors have contributed equally to this work.