Expanded Alphabet, Precise Sequencing Make DNA the Next Data Storage Solution

2022-03-06
Adding seven new letters to DNA’s molecular alphabet and developing a precise readout method enabled Illinois researchers to transform the double helix into a robust, sustainable data storage platform fit for the Information Age.

Imagine Bach’s “Cello Suite No. 1” played on a strand of DNA.

This scenario is not as impossible as it seems. Too small to withstand a rhythmic strum or sliding bowstring, DNA is a powerhouse for storing audio files and all kinds of other media.

“DNA is nature’s original data storage system. We can use it to store any kind of data: images, video, music — anything,” said Kasra Tabatabaei, a researcher at the Beckman Institute for Advanced Science and Technology and a coauthor on this study.

Expanding DNA’s molecular makeup and developing a precise new sequencing method enabled a multi-institutional team to transform the double helix into a robust, sustainable data storage platform.

The team's paper appeared in Nano Letters in February 2022.

In the age of digital information, anyone brave enough to navigate the daily news feels the global archive growing heavier by the day. Increasingly, paper files are being digitized to save space and protect information from natural disasters.

From scientists to social media influencers, anyone with information to store stands to benefit

from a secure, sustainable data lock box — and the double helix fits the bill.

“DNA is one of the best options, if not the best option, to store archival data especially,” said Chao Pan, a graduate student at the University of Illinois Urbana-Champaign and a coauthor on this study.

Its longevity rivaled only by durability, DNA is designed to weather Earth’s harshest conditions — sometimes for tens of thousands of years — and remain a viable data source. Scientists can sequence fossilized strands to uncover genetic histories and breathe life into long-lost landscapes.

Despite its diminutive stature, DNA is a bit like Dr. Who’s infamous police box: bigger on the inside than it appears.

“Every day, several petabytes of data are generated on the internet. Only one gram of DNA would be sufficient to store that data. That’s how dense DNA is as a storage medium,” said Tabatabaei, who is also a fifth-year Ph.D. student.

Another important aspect of DNA is its natural abundance and near-infinite renewability, a trait not shared by the most advanced data storage system on the market today: silicon microchips, which often circulate for just decades before an unceremonious burial in a heap of landfilled e-waste.

“At a time when we are facing unprecedented climate challenges, the importance of sustainable storage technologies cannot be overestimated. New, green technologies for DNA recording are emerging that will make molecular storage even more important in the future,” said Olgica Milenkovic, the Franklin W. Woeltge Professor of Electrical and Computer Engineering and a co-PI on the study.

Envisioning the future of data storage, the interdisciplinary team examined DNA’s millennia-

old MO. Then, the researchers added their own 21st-century twist.

In nature, every strand of DNA contains four chemicals — adenine, guanine, cytosine, and thymine — often referred to by the initials A, G, C, and T. They arrange and rearrange themselves along the double helix into combinations that scientists can decode, or sequence, to make meaning.

The researchers expanded DNA's already broad capacity for information storage by adding seven synthetic nucleobases to the existing four-letter lineup.

"Imagine the English alphabet. If you only had four letters to use, you could only create so many words. If you had the full alphabet, you could produce limitless word combinations. That's the same with DNA. Instead of converting zeroes and ones to A, G, C, and T, we can convert zeroes and ones to A, G, C, T, and the seven new letters in the storage alphabet," Tabatabaei said.

Because this team is the first to use chemically modified nucleotides for information storage in DNA, members innovated around a unique challenge: not all current technology is capable of interpreting chemically modified DNA strands. To solve this problem, they combined machine learning and artificial intelligence to develop a first-of-its-kind DNA sequence readout processing method.

Their solution can discern modified chemicals from natural ones, and differentiate each of the seven new molecules from one another.

"We tried 77 different combinations of the 11 nucleotides, and our method was able to differentiate each of them perfectly," Pan said. "The deep learning framework as part of our method to identify different nucleotides is universal, which enables the generalizability of our approach to many other applications."

This letter-perfect translation comes courtesy of nanopores: proteins with an opening in the middle through which a DNA strand can easily pass. Remarkably, the team found that nanopores can detect and distinguish each individual monomer unit along the DNA strand — whether the units have natural or chemical origins.

“This work provides an exciting proof-of-principle demonstration of extending macromolecular data storage to non-natural chemistries, which hold the potential to drastically increase storage density in non-traditional storage media,” said Charles Schroeder, the James Economy Professor of Materials Science and Engineering and a co-PI on this study.

DNA literally made history by storing genetic information. By the looks of this study, the future of data storage is just as double-helical.

Read the original article on University of Illinois Urbana-Champaign.