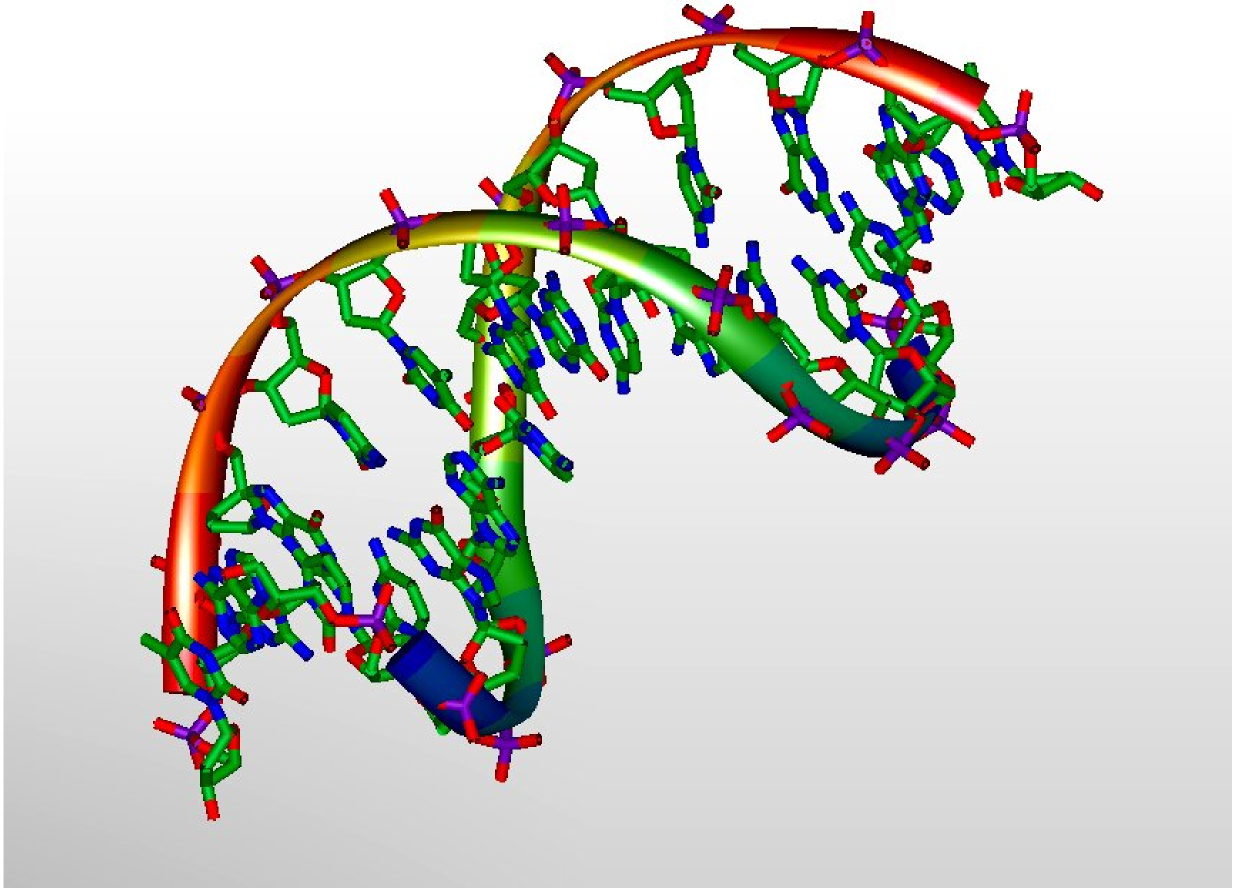
A technique for labeling and retrieving DNA data files from a large pool could help make DNA data storage feasible

June 10 2021



3D-model of DNA. Credit: Michael Ströck/Wikimedia/ GNU Free Documentation License

On Earth right now, there are about 10 trillion gigabytes of digital data, and every day, humans produce emails, photos, tweets, and other digital files that add up to another 2.5 million gigabytes of data. Much of this data is stored in enormous facilities known as exabyte data centers (an exabyte is 1 billion gigabytes), which can be the size of several football fields and cost around $1 billion to build and maintain.

Many scientists believe that an alternative solution lies in the molecule that contains our genetic information: DNA, which evolved to store massive quantities of information at very high density. A coffee mug full of DNA could theoretically store all of the world's data, says Mark Bathe, an MIT professor of biological engineering.

"We need new solutions for storing these massive amounts of data that the world is accumulating, especially the archival data," says Bathe, who is also an associate member of the Broad Institute of MIT and Harvard. "DNA is a thousandfold denser than even flash memory, and another property that's interesting is that once you make the DNA polymer, it doesn't consume any energy. You can write the DNA and then store it forever."

Scientists have already demonstrated that they can encode images and pages of text as DNA. However, an easy way to pick out the desired file from a mixture of many pieces of DNA will also be needed. Bathe and his colleagues have now demonstrated one way to do that, by encapsulating each data file into a 6-micrometer particle of silica, which is labeled with short DNA sequences that reveal the contents.

Using this approach, the researchers demonstrated that they could accurately pull out individual images stored as DNA sequences from a set of 20 images. Given the number of possible labels that could be used, this approach could scale up to 10^{20} files.

Bathe is the senior author of the study, which appears today in *Nature Materials*. The lead authors of the paper are MIT senior postdoc James Banal, former MIT research associate Tyson Shepherd, and MIT graduate student Joseph Berleant.

Stable storage

Digital storage systems encode text, photos, or any other kind of information as a series of 0s and 1s. This same information can be encoded in DNA using the four nucleotides that make up the genetic code: A, T, G, and C. For example, G and C could be used to represent 0 while A and T represent 1.

DNA has several other features that make it desirable as a storage medium: It is extremely stable, and it is fairly easy (but expensive) to synthesize and sequence. Also, because of its high density—each nucleotide, equivalent to up to two bits, is about 1 cubic nanometer—an exabyte of data stored as DNA could fit in the palm of your hand.

One obstacle to this kind of data storage is the cost of synthesizing such large amounts of DNA. Currently it would cost $1 trillion to write one petabyte of data (1 million gigabytes). To become competitive with magnetic tape, which is often used to store archival data, Bathe estimates that the cost of DNA synthesis would need to drop by about six orders of magnitude. Bathe says he anticipates that will happen within a decade or two, similar to how the cost of storing information on flash drives has dropped dramatically over the past couple of decades.

Aside from the cost, the other major bottleneck in using DNA to store data is the difficulty in picking out the file you want from all the others.

"Assuming that the technologies for writing DNA get to a point where it's cost-effective to write an exabyte or zettabyte of data in DNA, then what? You're going to have a pile of DNA, which is a gazillion files, images or movies and other stuff, and you need to find the one picture or movie you're looking for," Bathe says. "It's like trying to find a needle in a haystack."

Currently, DNA files are conventionally retrieved using PCR (polymerase chain reaction). Each DNA data file includes a sequence that binds to a particular PCR primer. To pull out a specific file, that primer is added to the sample to find and amplify the desired sequence. However, one drawback to this approach is that there can be crosstalk between the primer and off-target DNA sequences, leading unwanted files to be pulled out. Also, the PCR retrieval process requires enzymes and ends up consuming most of the DNA that was in the pool.

"You're kind of burning the haystack to find the needle, because all the other DNA is not getting amplified and you're basically throwing it away," Bathe says.

File retrieval

As an alternative approach, the MIT team developed a new retrieval technique that involves encapsulating each DNA file into a small silica particle. Each capsule is labeled with single-stranded DNA "barcodes" that correspond to the contents of the file. To demonstrate this approach in a cost-effective manner, the researchers encoded 20 different images into pieces of DNA about 3,000 nucleotides long, which is equivalent to about 100 bytes. (They also showed that the capsules could fit DNA files up to a gigabyte in size.)

Each file was labeled with barcodes corresponding to labels such as "cat" or "airplane." When the researchers want to pull out a specific image, they remove a sample of the DNA and add primers that correspond to the labels they're looking for—for example, "cat," "orange," and "wild" for an image of a tiger, or "cat," "orange," and "domestic" for a housecat.

The primers are labeled with fluorescent or magnetic particles, making it easy to pull out and identify any matches from the sample. This allows the desired file to be removed while leaving the rest of the DNA intact to be put back into storage. Their retrieval process allows Boolean logic statements such as "president AND 18th century" to generate George Washington as a result, similar to what is retrieved with a Google image search.

"At the current state of our proof-of-concept, we're at the 1 kilobyte per second search rate. Our file system's search rate is determined by the data size per capsule, which is currently limited by the prohibitive cost to write even 100 megabytes worth of data on DNA, and the number of sorters we can use in parallel. If DNA synthesis becomes cheap enough, we would be able to maximize the data size we can store per file with our approach," Banal says.

For their barcodes, the researchers used single-stranded DNA sequences from a library of 100,000 sequences, each about 25 nucleotides long, developed by Stephen Elledge, a professor of genetics and medicine at Harvard Medical School. If you put two of these labels on each file, you can uniquely label 10^{10} (10 billion) different files, and with four labels on each, you can uniquely label 10^{20} files.

Bathe envisions that this kind of DNA encapsulation could be useful for storing "cold" data, that is, data that is kept in an archive and not accessed very often. His lab is spinning out a startup, Cache DNA, that is now developing technology for long-term storage of DNA, both for DNA data storage in the long-term, and clinical and other preexisting DNA samples in the near-term.

"While it may be a while before DNA is viable as a data storage medium, there already exists a pressing need today for low-cost, massive storage solutions for preexisting DNA and RNA samples from COVID-19 testing, human genomic sequencing, and other areas of genomics," Bathe says.

More information: Random access DNA memory using Boolean search in an archival file storage system, *Nature Materials* (2021). DOI: 10.1038/s41563-021-01021-3 , www.nature.com/articles/s41563-021-01021-3

Provided by Massachusetts Institute of Technology

Citation: A technique for labeling and retrieving DNA data files from a large pool could help make DNA data storage feasible (2021, June 10) retrieved 23 April 2024 from https://phys.org/news/2021-06-technique-dna-large-pool-storage.html