Los Alamos
NATIONAL LABORATORY
EST 1943

Translation software enables efficient DNA data storage

April 1, 2021

LOS ALAMOS, N.M., April 1, 2021—In support of a major collaborative project to store massive amounts of data in DNA molecules, a Los Alamos National Laboratory–led team has developed a key enabling technology that translates digital binary files into the four-letter genetic alphabet needed for molecular storage.

“Our software, the Adaptive DNA Storage Codec (ADS Codex), translates data files from what a computer understands into what biology understands,” said Latchesar Ionkov, a computer scientist at Los Alamos and principal investigator on the project. “It’s like translating from English to Chinese, only harder.”

The work is key part of the Intelligence Advanced Research Projects Activity (IARPA) Molecular Information Storage (MIST) program to bring cheaper, bigger, longer-lasting storage to big-data operations in government and the private sector. The short-term goal of MIST is to write 1 terabyte—a trillion bytes—and read 10 terabytes within 24 hours for $1,000. Other teams are refining the writing (DNA synthesis) and retrieval (DNA sequencing) components of the initiative, while Los Alamos is working on coding and decoding.

“DNA offers a promising solution compared to tape, the prevailing method of cold storage, which is a technology dating to 1951,” said Bradley Settlemyer, a storage systems researcher and systems programmer specializing in high-performance computing at Los Alamos. “DNA storage could disrupt the way we think about archival storage, because the data retention is so long and the data density so high. You could store all of YouTube in your refrigerator, instead of in acres and acres of data centers. But researchers first have to clear a few daunting technological hurdles related to integrating different technologies.”

Not lost in translation

Compared to the traditional long-term storage method that uses pizza-sized reels of magnetic tape, DNA storage is potentially less expensive, far more physically compact, more energy efficient, and longer lasting—DNA survives for hundreds of years and doesn’t require maintenance. Files stored in DNA also can be very easily copied for negligible cost.

DNA’s storage density is staggering. Consider this: humanity will generate an estimated 33 zettabytes by 2025—that’s 3.3 followed by 22 zeroes. All that information would fit into a ping pong ball, with room to spare. The Library of Congress has about 74 terabytes, or 74 million million bytes, of information—6,000 such libraries would fit in a DNA archive the size of a poppy seed. Facebook’s 300 petabytes (300,000 terabytes) could be stored in a half poppy seed.

Encoding a binary file into a molecule is done by DNA synthesis. A fairly well understood technology, synthesis organizes the building blocks of DNA into various arrangements, which are indicated by sequences of the letters A, C, G, and T. They are the basis of all DNA code, providing the instructions for building every living thing on earth.

The Los Alamos team’s ADS Codex tells exactly how to translate the binary data—all 0s and 1s—into sequences of four letter-combinations of A, C, G, and T. The Codex also handles the decoding back into binary. DNA can be synthesized by several methods, and ADS Codex can accommodate them all. The Los Alamos team has completed a version 1.0 of ADS Codex and in November 2021 plans to use it to evaluate the storage and retrieval systems developed by the other MIST teams.

Unfortunately, DNA synthesis sometimes makes mistakes in the coding, so ADS Codex addresses two big obstacles to creating DNA data files.

First, compared to traditional digital systems, the error rates while writing to molecular storage are very high, so the team had to figure out new strategies for error correction. Second, errors in DNA storage arise from a different source than they do in the digital world, making the errors trickier to correct.

“On a digital hard disk, binary errors occur when a 0 flips to a 1, or vice versa, but with DNA, you have more problems that come from insertion and deletion errors,” Ionkov said. “You’re writing A, C, G, and T, but sometimes you try to write A, and nothing appears, so the sequence of letters shifts to the left, or it types AAA. Normal error correction codes don’t work well with that.”

ADS Codex adds additional information called error detection codes that can be used to validate the data. When the software converts the data back to binary, it tests if the codes match. If they don’t, ACOMA tries removing or adding nucleotides until the verification succeeds.

Smart scale-up

Large warehouses contain today’s largest data centers, with storage at the exabyte scale—that’s a trillion million bytes or more. Costing billions to build, power, and run, this type of digitally based data centers may not be the best option as the need for data storage continues to grow exponentially.

Long-term storage with cheaper media is important for the national security mission of Los Alamos and others. “At Los Alamos, we have some of the oldest digital-only data and largest stores of data, starting from the 1940s,” Settlemyer said. “It still has tremendous value. Because we keep data forever, we’ve been at the tip of the spear for a long time when it comes to finding a cold-storage solution.”

Settlemyer said DNA storage has the potential to be a disruptive technology because it crosses between fields ripe with innovation. The MIST project is stimulating a new coalition among legacy storage vendors who make tape, DNA synthesis companies, DNA sequencing companies, and high-performance computing organizations like Los Alamos that are driving computers into ever-larger-scale regimes of science-based simulations that yield mind-boggling amounts of data that must be analyzed.

Deeper dive into DNA

When most people think of DNA, they think of life, not computers. But DNA is itself a four-letter code for passing along information about an organism. DNA molecules are made from four types of bases, or nucleotides, each identified by a letter: adenine (A), thymine (T), guanine (G), and cytosine (C).

These bases wrap in a twisted chain around each other—the familiar double helix—to form the molecule. The arrangement of these letters into sequences creates a code that tells an organism how to form. The complete set of DNA molecules makes up the genome—the blueprint of your body.

By synthesizing DNA molecules—making them from scratch—researchers have found they can specify, or write, long strings of the letters A, C, G, and T and then read those sequences back. The process is analogous to how a computer stores information using 0s and 1s. The method has been proven to work, but reading and writing the DNA-encoded files currently takes a long time, Ionkov said.

“Appending a single nucleotide to DNA is very slow. It takes a minute,” Ionkov said. “Imagine writing a file to a hard drive taking more than a decade. So that problem is solved by going massively parallel. You write tens of millions of molecules simultaneously to speed it up.”

While various companies are working on different ways of synthesizing to address this problem, ADS Codex can be adapted to every approach.

Funding: Funding for ADS Codex was provided by the Intelligence Advanced Research Projects Activity (IARPA), a research agency within the Office of the Director of National Intelligence.