All of YouTube in a single teaspoon: Storing information in DNA

September 11 2019, by Kevin Hattori

A depiction of the double helical structure of DNA. Its four coding units (A, T, C, G) are color-coded in pink, orange, purple and yellow. Credit: NHGRI

Researchers at the Technion–Israel Institute of Technology in Haifa and the Interdisciplinary Center (IDC) Herzliya have demonstrated a significant improvement in the efficiency of the process needed to store digital information in DNA.

In a paper published in *Nature Biotechnology*, the group demonstrated storage of information in a density of more than 10 petabytes (10 million gigabytes) in a single gram of DNA while significantly improving the writing process. Theoretically, such density allows for storing all the information stored on YouTube in a volume of a single teaspoon.

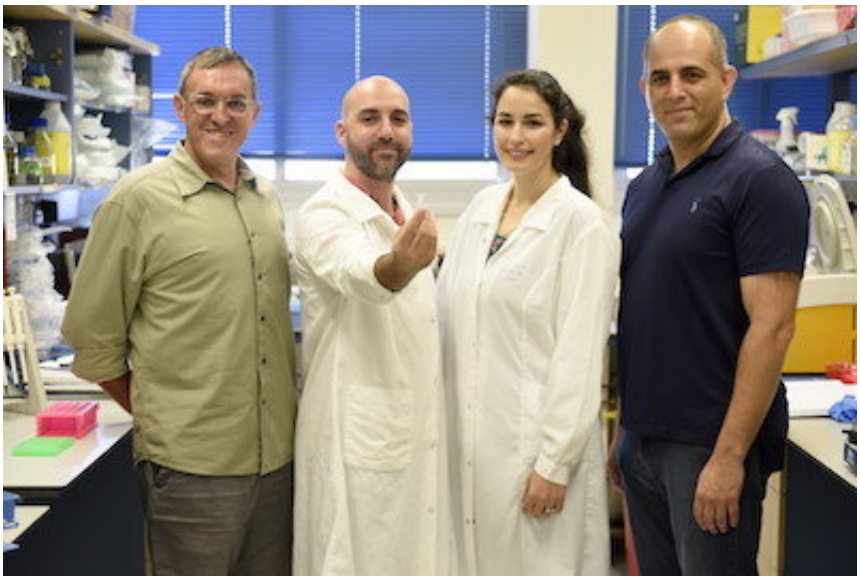
The study was led by research student Leon Anavy, a student in the Technion Faculty of Computer Science, under the guidance of Professor Zohar Yakhini of the Technion Faculty of Computer Science and the Efi Arazi School of Computer Science at the Interdisciplinary Center Herzliya. It was conducted in collaboration with Professor Roee Amit's Synthetic Biology Laboratory at the Technion Faculty of Biotechnology and Food Engineering.

The amount of digital information available to humanity has grown at a tremendous speed since IBM invented the hard disk in the 1950s. Storing this information has become a major challenge not only in the technological context but also with regards to economic and environmental aspects, as server farms—information warehouses that serve us all—are currently responsible for about 2 percent of global carbon emissions, a similar rate to the cumulative emission of global air traffic, and for about 3 percent of global electricity consumption, more than the electricity consumption of the entire UK. Against this backdrop, a new technological approach has developed over the last decade: information storage in DNA. This technology allows for significant minimization, longer-term (thousand-fold) retention of information, and zero energy and economic cost of maintenance.

The basic idea of encoding information on DNA is that the DNA molecule is a chain made up of links called nucleotides. The nucleotides are divided into four types marked with letters A, C, G and T. To store information on DNA, each binary sequence (consisting of the 0 and 1 symbols) must be translated into a sequence consisting of these letters. In the next step, in a process called synthesis, actual DNA molecules are produced representing these same sequences. To read the data, these DNA molecules are sequenced. DNA sequencing produces an output that represents the nucleotide sequence that makes up each molecule in the input. That output is then translated into a binary sequence that represents the original message that was coded. Modern technologies support the synthesis of many thousands of different nucleotide series in parallel.

The storage of information on DNA is a very complex technological challenge. In the field of information reading (sequencing), there has been tremendous progress driven by the genome revolution; for the writing of information, however, there are still significant technological difficulties and costs are heavier. This is the importance of the researchers' breakthrough. It allows for: (1) increasing the number of letters used to encode the information (beyond the original 4 letters); (2) significantly reducing the number of synthesis rounds required to store information on DNA; (3) improving the error correction mechanism used.

Researchers at the Technion and at IDC Herzliya have increased the effective number of letters beyond the four building blocks in natural DNA, using new letters that are unique combinations of the original letters. The idea is similar to the formation of new colors using mixtures of base colors. Increasing the number of letters allows more information to be encoded in each letter in the sequence.



From left: Prof. Zohar Yakhini, Leon Anavy, Inbal Vaknin, and Prof. Roee Amit. Credit: American Technion Society

According to Prof. Yakhini, "The current synthesis and sequencing processes are inherently redundant, because each molecule is produced in large numbers1 and is read in multiple copies during sequencing. The method we developed leverages this redundancy to increase the effective number of letters well over the original four letters, making it possible for us to encode and write each unit of information in fewer cycles of synthesis."

The team demonstrated a reduction of the number of synthesis rounds required per unit of information by 20 percent. They also showed that the number of synthesis rounds could be reduce in the future by 75 percent without significant development efforts. This means that the storage process will be faster and less expensive.

"In this work, we have implemented a DNA based storage system that encodes information with synthesis efficiency that is significantly better

than the standard approach," explained Prof. Amit. "The study included the actual implementation of the new coding technique for storing large-volume information on DNA molecules and reconstructing it for testing the process."

In fact, on one of the shelves in Prof. Amit's lab at the Technion sits a small test tube containing about 10 nanograms (billionths of a gram) of DNA, encoding thousands of copies of a bilingual version of the Bible.

The research group has developed advanced error correction mechanisms to overcome errors that are an integral part of biological-physical processes, like the one used here. Part of the DNA sequence of the molecules that store the information, designed by Leon Anavy and Prof. Yakhini, is used for this error correction.

According to Leon Anavy, "thanks to the use of error-correction codes that are tailored to the unique encoding we created, we were able to perform highly efficient coding and to successfully recover the information. When working in a system consisting of millions of parts (molecules), even one-in-a-million events occur, which can disrupt the reading. Careful coding allowed us to overcome these problems."

According to the researchers, "the technology we presented in the paper has the potential to streamline further processes in synthetic biology and biotechnology. We believe that in the coming years, we will see a significant increase in the use of synthetic DNA in research and industry."

The synthetic DNA used by the researchers and designed by the group was produced by the Twist Bioscience, a California based company that also has offices in Tel Aviv. Sequencing was performed at the Technion's Genome Center. The study was partly supported by the European Commission's Horizon 2020 Framework Program for

Research and Innovation. Leon Anavy is a fellow of the ADAMS fellowship program of the Israeli Science Academy. Dr. Orna Atar and research student Inbal Vaknin were also involved in the study.

More information: Leon Anavy et al. Data storage in DNA with fewer synthesis cycles using composite DNA letters, *Nature Biotechnology* (2019). DOI: 10.1038/s41587-019-0240-x

Provided by American Technion Society