Researchers Store Computer Operating System and Short Movie on DNA

New Coding Strategy Maximizes Data Storage Capacity of DNA Molecules

New York, NY - March 2, 2017 -- Humanity may soon generate more data than hard drives or magnetic tape can handle, a problem that has scientists turning to nature's age-old solution for information-storage — DNA.

In a new study in *Science*, a pair of researchers at Columbia University and the New York Genome Center (NYGC) show that an algorithm designed for streaming video on a cellphone can unlock DNA's nearly full storage potential by squeezing more information into its four base nucleotides. They demonstrate that this technology is also extremely reliable.

DNA is an ideal storage medium because it's ultra-compact and can last hundreds of thousands of years if kept in a cool, dry place, as demonstrated by the recent recovery of DNA from the bones of a 430,000-year-old human ancestor found in a cave in Spain.

"DNA won't degrade over time like cassette tapes and CDs, and it won't become obsolete—if it does, we have bigger problems," said study coauthor Yaniv Erlich, a computer science professor at Columbia Engineering, a member of Columbia's Data Science Institute, and a core member of the NYGC.

Erlich and his colleague Dina Zielinski, an associate scientist at NYGC, chose six files to encode, or write, into DNA: a full computer operating system, an 1895 French film, "Arrival of a train at La Ciotat," a $50 Amazon gift card, a computer virus, a Pioneer plaque and a 1948 study by information theorist Claude Shannon.

They compressed the files into a master file, and then split the data into short strings of binary code made up of ones and zeros. Using an erasure-correcting algorithm called fountain codes, they randomly packaged the strings into so-called droplets, and mapped the ones and zeros in each droplet to the four nucleotide bases in DNA: A, G, C and T. The algorithm deleted letter combinations known to create errors, and added a barcode to each droplet to help reassemble the files later.

In all, they generated a digital list of 72,000 DNA strands, each 200 bases long, and sent it in a text file to a San Francisco DNA-synthesis startup, Twist Bioscience, that specializes in turning digital data into biological data. Two weeks later, they received a vial holding a speck of DNA molecules.

To retrieve their files, they used modern sequencing technology to read the DNA strands, followed by software to translate the genetic code back into binary. They recovered their files with zero errors, the study reports. (In this short demo, Erlich opens his archived operating system on a virtual machine and plays a game of Minesweeper to celebrate.)

They also demonstrated that a virtually unlimited number of copies of the files could be created with their coding technique by multiplying their DNA sample through polymerase chain reaction (PCR), and that those copies, and even copies of their copies, and so on, could be recovered error-free.

Finally, the researchers show that their coding strategy packs 215 petabytes of data on a single gram of DNA—100 times more than methods published by pioneering researchers George Church at Harvard, and Nick Goldman and Ewan Birney at the European Bioinformatics Institute. “We believe this is the highest-density data-storage device ever created,” said Erlich.

The capacity of DNA data-storage is theoretically limited to two binary digits for each nucleotide, but the biological constraints of DNA itself and the need to include redundant information to reassemble and read the fragments later reduces its capacity to 1.8 binary digits per nucleotide base.

The team’s insight was to apply fountain codes, a technique Erlich remembered from graduate school, to make the reading and writing process more efficient. With their DNA Fountain technique, Erlich and Zielinski pack an average of 1.6 bits into each base nucleotide. That’s at least 60 percent more data than previously published methods, and close to the 1.8-bit limit.

Cost still remains a barrier. The researchers spent $7,000 to synthesize the DNA they used to archive their 2 megabytes of data, and another $2,000 to read it. Though the price of DNA sequencing has fallen exponentially, there may not be the same demand for DNA synthesis, says Sri Kosuri, a biochemistry professor at UCLA who was not involved in the study. “Investors may not be willing to risk tons of money to bring costs down,” he said.

But the price of DNA synthesis can be vastly reduced if lower-quality molecules are produced, and coding strategies like DNA Fountain are used to fix molecular errors, says Erlich. “We can do more of the heavy lifting on the computer to take the burden off time-intensive molecular coding,” he said.

###

Media contact:
New York Genome Center
Karen Zipern
(c): 917-415-8134 (o): 646-977-7065
kzipern@nygenome.org

Columbia
Kim Martineau
(o): 646-717-0134
klm32@columbia.edu

Scientist contacts:
Yaniv Erlich
617-913-1318
yaniv@columbia.cs.edu

Dina Zielinski
dzielinksi@nygenome.org

The New York Genome Center

The New York Genome Center is an independent, nonprofit academic research organization at the forefront of transforming biomedical research and clinical care with the mission of saving lives. A collaboration of renowned academic, medical and industry leaders across the globe, the New York Genome Center's goal is to translate genomic research into development of new treatments, therapies and therapeutics against human disease. Its member organizations and partners are united in this unprecedented collaboration of technology, science and medicine, designed to harness the power of innovation and discoveries to advance genomic services. The New York Genome Center's member organizations are united in this unprecedented collaboration of technology, science and medicine. Their shared objective is the acceleration of medical genomics and precision medicine to benefit patients around the world.

Member institutions include: Albert Einstein College of Medicine, American Museum of Natural History, Cold Spring Harbor Laboratory, Columbia University, Weill Cornell Medicine, Hospital for Special Surgery, The Jackson Laboratory, Memorial Sloan Kettering Cancer Center, Icahn School of Medicine at Mount Sinai, NewYork-Presbyterian Hospital, The New York Stem Cell Foundation, New York University, Northwell Health (formerly North Shore-LIJ), Princeton University, The Rockefeller University, Roswell Park Cancer Institute, Stony Brook University and IBM.