Researchers achieve random access in large-scale DNA data storage

February 21 2018

Allen School Ph.D. student Lee Organick (foreground) and Microsoft researcher Yuan-Jyue Chen in the Molecular Information Systems Lab. Credit: Dennis Wise/University of Washington

University of Washington and Microsoft researchers revealed today that they have taken a significant step forward in their quest to develop a DNA-based storage system for digital data. In a paper published in *Nature Biotechnology*, the members of the Molecular Information Systems Laboratory (MISL) describe the science behind their world record-setting achievement of 200 megabytes stored in synthetic DNA. They also present their system for random access—that is, the selective retrieval of individual data files encoded in more than 13 million DNA oligonucleotides. While this is not the first time researchers have achieved random access in DNA, the UW and Microsoft team have produced the first demonstration of random access at such a large scale.

One of the big advantages to DNA as a digital storage medium is its ability to store vast quantities of information, with a raw limit of one exabyte—equivalent to one billion gigabytes—per cubic millimeter. The data must be converted from digital 0s and 1s to the molecules of DNA: adenine, thymine, cytosine, and guanine. To restore the data to its digital form, the DNA is sequenced and the files decoded back to 0s and 1s. This process becomes more daunting as the amount of data increases—without the ability to perform random access, the entire dataset would have to be sequenced and decoded in bulk in order to find and retrieve specific files. In addition, the DNA synthesis and sequencing processes are error-prone, which can result in data loss.

MISL researchers addressed these problems by designing and validating an extensive library of primers for use in conjunction with polymerase chain reaction (PCR) to achieve random access. Before synthesizing the

DNA containing data from a file, the researchers appended both ends of each DNA sequence with PCR primer targets from the primer library. They then used these primers later to select the desired strands through random access, and used a new algorithm designed to more efficiently decode and restore the data to its original, digital state.

"Our work reduces the effort, both in sequencing capacity and in processing, to completely recover information stored in DNA," explained Microsoft Senior Researcher Sergey Yekhanin, who was instrumental in creating the codec and algorithms used to achieve the team's results. "For the latter, we have devised new algorithms that are more tolerant to errors in writing and reading DNA sequences to minimize the effort in recovering this information."

Using synthetic DNA supplied by Twist Bioscience, the MISL team encoded and successfully retrieved 35 distinct files ranging in size from 29 kilobytes to over 44 megabytes—amounting to a record-setting 200 megabytes of high-definition video, audio, images, and text. This represents a significant increase over the previous record of 22 megabytes set by researchers from Harvard Medical School and Technicolor Research & Innovation in Germany.

"The intersection of biotech and computer architecture is incredibly promising and we are excited to detail our results to the community," said Allen School professor Luis Ceze, who co-leads the MISL. "Since this paper was submitted for publication we have reached over 400 megabytes, and we are still growing and learning more about large-scale DNA data storage."

With this new milestone, MISL researchers have succeeded in demonstrating how DNA-based data storage—known to be significantly denser and more durable than existing digital storage technologies—can be practical, too. The UW and Microsoft team estimates its approach

will scale to physically isolated pools of DNA containing several terabytes each. When dehydrated for storage, these pools of data would be several orders of magnitude denser than tape. And as the costs associated with DNA sequencing and synthesis continue to decline, the team foresees substantial activity devoted to the development of DNA-based data storage in future.

"DNA data storage is an incredibly exciting area, and it is great to see our progress recognized by such a reputable publication as *Nature Biotechnology*," said Microsoft Senior Researcher Karin Strauss, co-leader of the MISL and an affiliate professor at the Allen School. "We are enthusiastic to continue working at the intersection of biotechnology and IT."

It was this intersection that initially interested Allen School Ph.D. student Lee Organick, who performed many of the wet-lab experiments the team used to validate its approach. Having made the leap from undergraduate studies in molecular biology to computer science, she is enthusiastic about the potential impact of the MISL's approach.

"We're at a time when a lot of groundbreaking research will be done at the intersection of fields," said Organick. "When I heard about this project it seemed a bit outlandish, but it captured my imagination."

The makeup of the lab—which unites researchers from multiple disciplines and organizations—is another plus, in Organick's view.

"Having worked with such a creative and diverse team of people for several years now, they've shown me that projects like this one are achievable," she said. "And it's just as exciting as it was the first day."

More information: Lee Organick et al. Random access in large-scale DNA data storage, *Nature Biotechnology* (2018). DOI: 10.1038/nbt.4079

DNA Data Storage Gets Random Access. spectrum.ieee.org/the-human-os … e-gets-random-access

Provided by University of Washington