April 7, 2016

UW team stores digital images in DNA — and retrieves them perfectly

Jennifer Langston
News and Information

Technology companies routinely build sprawling data centers to store all the baby pictures, financial transactions, funny cat videos and email messages its users hoard.

All the movies, images, emails and other digital data from more than 600 basic smartphones (10,000 gigabytes) can be stored in the faint pink smear of DNA at the end of this test tube. ***Tara Brown Photography/ University of Washington***

But a new technique developed by University of Washington and Microsoft researchers could shrink the space needed to store digital data that today would fill a Walmart supercenter down to the size of a sugar cube.

For more information, visit the Molecular Information Systems Lab.

In a paper presented in April at the ACM International Conference on Architectural Support for Programming Languages and Operating Systems, the team of computer scientists and electrical engineers has detailed one of the first complete systems to encode, store and retrieve digital data using DNA molecules, which can store information millions of times more compactly than current archival technologies.

Authors of the paper are UW computer science and engineering doctoral student James Bornholt, UW bioengineering doctoral student Randolph Lopez, UW associate professor of computer science and engineering Luis Ceze, UW associate professor of electrical engineering and of computer science and engineering Georg Seelig, and Microsoft researchers and UW CSE affiliate faculty Doug Carmean and Karin Strauss.

In one experiment outlined in the paper, the team successfully encoded digital data from four image files into the nucleotide sequences of synthetic DNA snippets. More significantly, they were also able to reverse that process — retrieving the correct sequences from a larger pool of DNA and reconstructing the images without losing a single byte of information.

The team has also encoded and retrieved data that authenticates archival video files from the UW's Voices from the Rwanda Tribunal project that contain interviews with judges, lawyers and other personnel from the Rwandan war crime tribunal.

"Life has produced this fantastic molecule called DNA that efficiently stores all kinds of information about your genes and how a living system works — it's very, very compact and very durable," said co-author Luis Ceze, UW associate professor of computer science and engineering.

"We're essentially repurposing it to store digital data — pictures, videos, documents — in a manageable way for hundreds or thousands of years."

Lee Organick, a UW computer science and engineering research scientist, mixes DNA samples for storage. Each tube contains a digital file, which might be a picture of a cat or a Tchaikovsky symphony. ***Tara Brown Photography/ University of Washington***

The digital universe — all the data contained in our computer files, historic archives, movies, photo collections and the exploding volume of digital information collected by businesses and devices worldwide — is expected to hit 44 trillion gigabytes by 2020.

That's a tenfold increase compared to 2013, and will represent enough data to fill more than six stacks of computer tablets stretching to the moon. While not all of that information needs to be saved, the world is producing data faster than the capacity to store it.

DNA molecules can store information many millions of times more densely than existing technologies for digital storage — flash drives, hard drives, magnetic and optical media. Those systems also degrade after a few years or decades, while DNA can reliably preserve information for centuries. DNA is best suited for archival applications, rather than instances where files need to be accessed immediately.

The team from the Molecular Information Systems Lab housed in the UW Electrical Engineering Building, in close collaboration with Microsoft Research, is developing a DNA-based storage system that it expects could address the world's needs for archival storage.

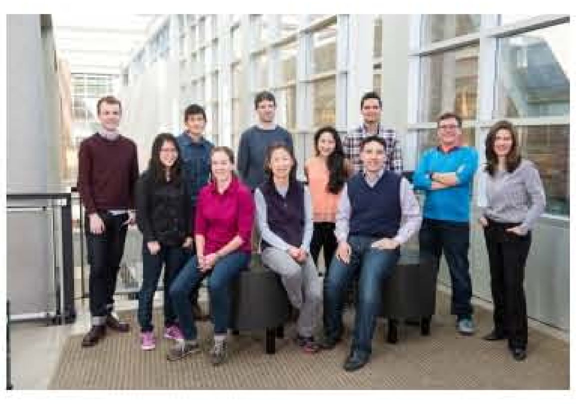
First, the researchers developed a novel approach to convert the long strings of ones and zeroes in digital data into the four basic building blocks of DNA sequences — adenine, guanine, cytosine and thymine.

"How you go from ones and zeroes to As, Gs, Cs and Ts really matters because if you use a smart approach, you can make it very dense and you don't get a lot of errors," said co-author Georg Seelig, a UW associate professor of electrical engineering and of computer science and engineering. "If you do it wrong, you get a lot of mistakes."

The digital data is chopped into pieces and stored by synthesizing a massive number of tiny DNA molecules, which can be dehydrated or otherwise preserved for long-term storage.

The UW and Microsoft researchers are one of two teams nationwide that have also demonstrated the ability to perform "random access" — to identify and retrieve the correct sequences from this large pool of random DNA molecules, which is a task similar to reassembling one chapter of a story from a library of torn books.

To access the stored data later, the researchers also encode the equivalent of zip codes and street addresses into the DNA sequences. Using Polymerase Chain Reaction (PCR) techniques — commonly used in molecular biology — helps them more easily identify the zip codes they are looking for. Using DNA sequencing techniques, the researchers can then "read" the data and convert them back to a video, image or document file by using the street addresses to reorder the data.



The Molecular Information Systems Lab research team: Front (left to right): Bichlien Nguyen, Lee Organick, Hsing-Yeh Parker, Siena Dumas Ang, Chris Takahashi. Back (left to right): James Bornholt, Yuan-Jyue Chen, Georg Seelig, Randolph Lopez, Luis Ceze, Karin Strauss. Not pictured: Doug Carmean, Rob Carlson, Krittika d'Silva. Credit: Tara Brown Photography/University of Washington ***Tara Brown Photography/University of Washington***

Currently, the largest barrier to viable DNA storage is the cost and efficiency with which DNA can be synthesized (or manufactured) and sequenced (or read) on a large scale. But researchers say there's no technical barrier to achieving those gains if the right incentives are in place.

Advances in DNA storage rely on techniques pioneered by the biotechnology industry, but also incorporate new expertise. The team's encoding approach, for instance, borrows from error correction schemes commonly used in computer memory — which hadn't been applied to DNA.

"This is an example where we're borrowing something from nature — DNA — to store information. But we're using something we know from computers — how to correct memory errors — and applying that back to nature," said Ceze.

"This multidisciplinary approach is what makes this project exciting. We are drawing from a diverse set of disciplines to push the boundaries of what can be done with DNA. And, as a result, creating a storage system with unprecedented density and durability," said Karin Strauss, a researcher at Microsoft and UW affiliate associate professor of computer science and engineering.

The research was funded by Microsoft Research, the National Science Foundation, and the David Notkin Endowed Graduate Fellowship.

For more information, contact Ceze at luisceze@cs.washington.edu or Seelig at gseelig@u.washington.edu. To reach Strauss, please contact TNRPR@we-worldwide.com.

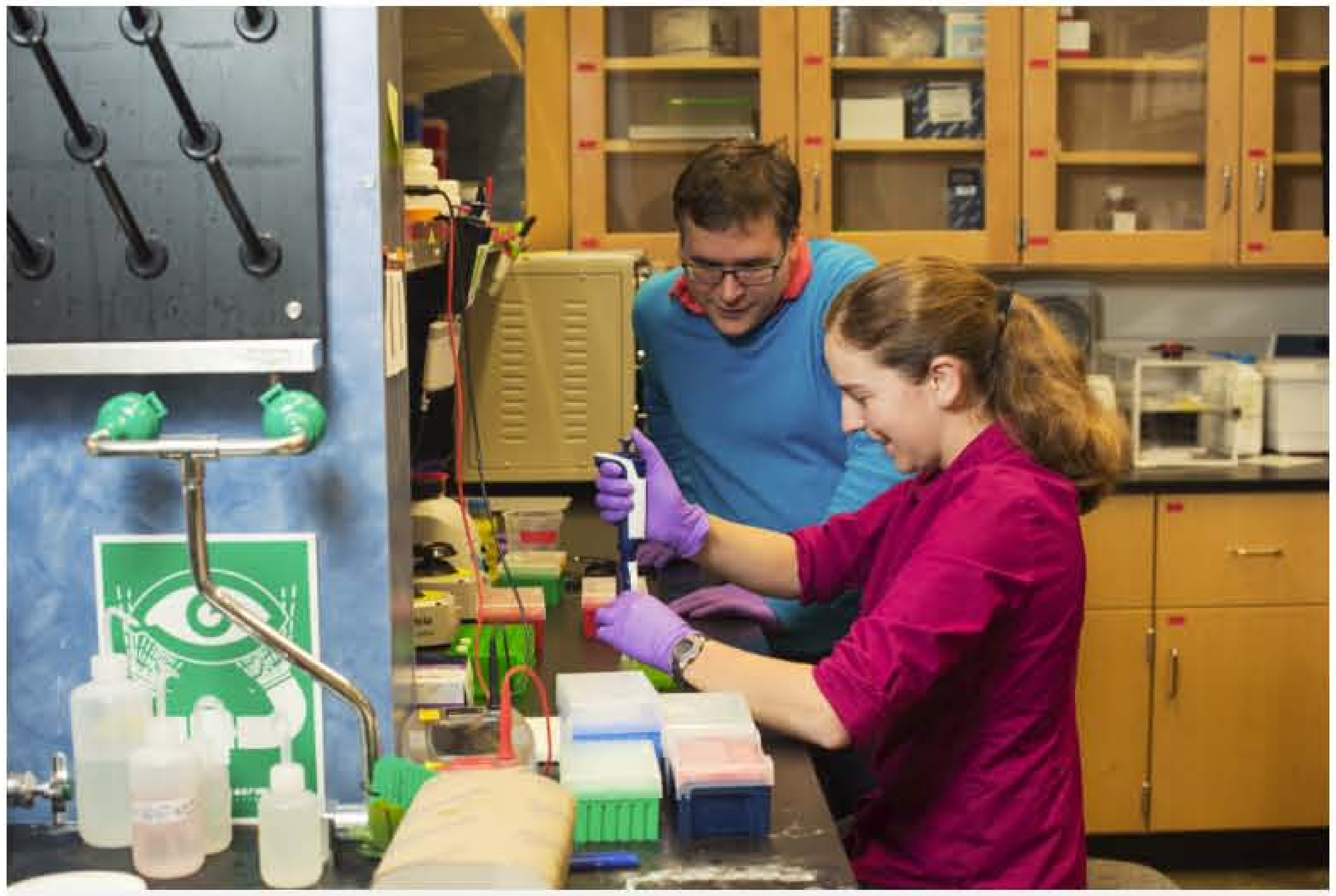
Tag(s): College of Engineering • Department of Computer Science & Engineering • Georg Seelig • Luis Ceze

04/29/2016

Plugging Into DNA for Digital Data Storage

6:34 minutes

LISTEN ON SOUNDCLOUD MORE FROM THIS EPISODE



Luis Ceze, the University of Washington Torode Family Career Development Professor of Computer Science & Engineering, and research scientist Lee Organick prepare DNA containing digital data for sequencing, which allows them to "read" and retrieve the original files. Credit: Tara Brown Photography/University of Washington

DNA is the storage system for the biological code of the human genome. Now, engineers are tapping into this natural code to store digital data. For instance, Georg Seelig, an engineer from the University of Washington, and his team were able to store and retrieve digitized photos on strands of DNA. Seelig discusses how to translate binary code into the four nucleotide bases.

Segment Guests

Georg Seelig

Georg Seelig is an associate professor in Electrical Engineering and Computer Science & Engineering at the University of Washington in Seattle, Washington.

MORE FROM GUEST

Researchers store, search, retrieve images in DNA

DNA is the future of data storage, and it just got more viable as a team of researchers find a way to encode direct access.

Sci-Tech

by **Michelle Starr**
April 11, 2016 12:37 AM PDT
@riding_red

A team at the University of Washington has moved DNA data storage forward a significant step by making the information both searchable and directly accessible.

The pink smear in the test tube could hold up to 10,000 gigabytes of data.

Tara Brown Photography/ University of Washington

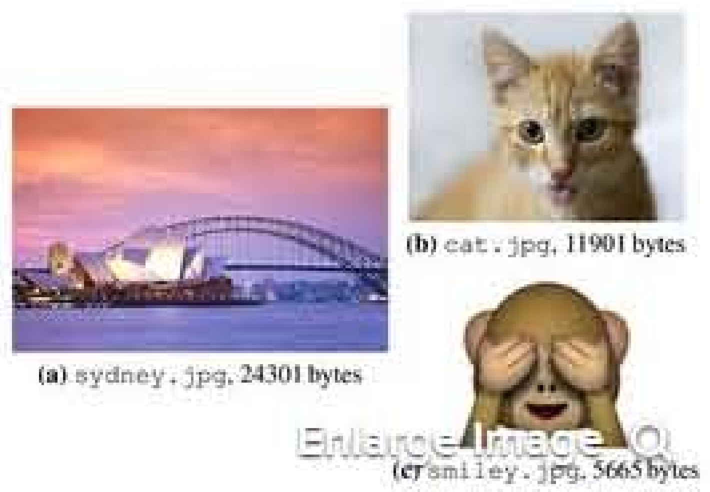
They encoded four digital images in DNA, and then retrieved them perfectly. A **paper** released last week details the effort.

The past few years have seen important strides in using DNA to store digital data. In 2012, Harvard researchers demonstrated that 5.5 petabits (5,500 terabits) of data can be stored in a single cubic millimetre of DNA. In 2013, researchers from the European Bioinformatics Institute showed that data could be retrieved by sequencing the DNA.

"Life has produced this fantastic molecule called DNA that efficiently stores all kinds of information about your genes and how a living system works -- it's very, very compact and very durable," paper co-author Luis Ceze, associate professor of computer science and engineering, said last week in a **statement**.

"We're essentially repurposing it to store digital data -- pictures, videos, documents -- in a **DNA**.

The other really cool part is the direct access (also known as random access) and searchability of the data encoded on the DNA, which eliminates the previous need to sequence the entire DNA to find the information.



Here are three of the images the team encoded into DNA.

University of Washington

"Suppose you have a large amount of information encoded in DNA in a big pool -- think petabytes," Ceze explained. "That information is stored in a large collection of small DNA molecules. How would you read just a small specific part of the data, say a video in a large video collection? Without random access you need to access the whole thing until you find what you want. With random access, you can access the desired data directly."

To achieve this, the team used something called **Huffman coding**, which is usually used in lossless data compression.

At the moment, DNA data storage is still prohibitively expensive, and the process of DNA synthesis is far from perfect. However, this combination of error correction and random access pushes us further into a future where the equivalent of a Walmart of digital data could be stored in a space the size of a sugar cube.

TECH & SCIENCE

DNA STORAGE COULD MAKE DATA CENTERS OBSOLETE

BY ANTHONY CUTHBERTSON ON 4/15/16 AT 3:33 PM

01:10 01:11 HD

TECH & SCIENCE DATA DNA DATA STORAGE

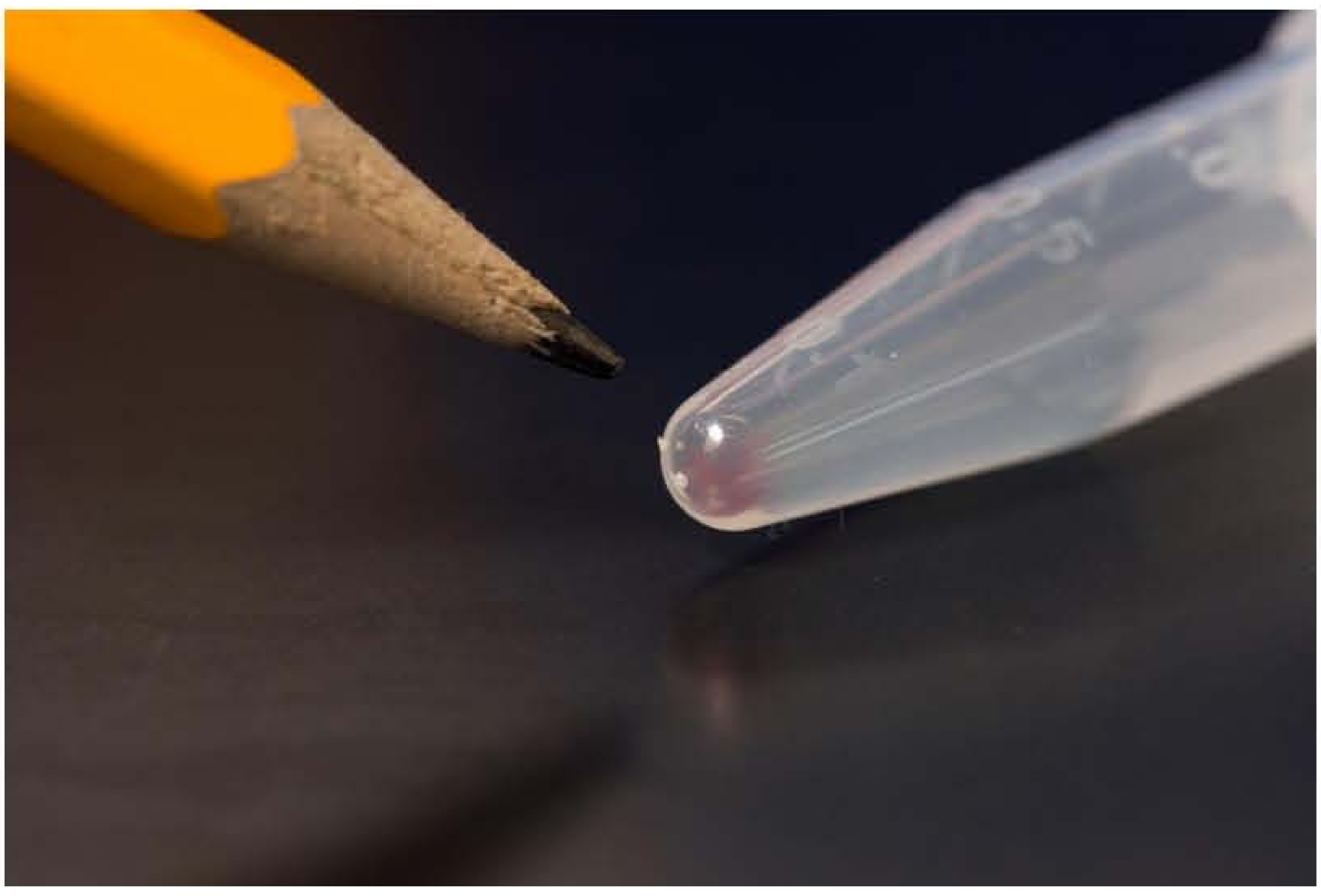
Imagine being able to store a data center's worth of information on something the size of a single USB stick. A new technique developed by a team of computer scientists could one day make this possible using something borrowed from nature: DNA.

All the data stored in a warehouse-sized center could fit in the size of a sugar cube using a new DNA data storage technique.

CREATIVE COMMONS

The images, which are encoded as a string of 0's and 1's, are converted into a string of A's, C's, T's and G's—the bases that pair to form DNA. A DNA molecule with that sequence is then chemically synthesized and dried out for storing with billions of other molecules.

"This works for any digital data, not just images," says Luis Ceze, an associate professor of electrical engineering at the University of Washington and one of the authors of the study. "We used images because images and video tend to take lots of storage space."



Around 10,000 gigabytes of data can be stored in the faint pink smear of DNA at the end of this test tube.

TARA BROWN PHOTOGRAPHY/ UNIVERSITY OF WASHINGTON

Digital data—including videos, photos and text—collected by devices is expected to hit 44 trillion gigabytes by 2020. At its current rate, the world is producing more data than the capacity to store it. Ceze and his colleagues believe DNA could be the solution to this problem.

Related Stories

- Plant Forensics Aid Crime Investigations
- Ultimate Time Capsule Stores Data 'For Eternity'
- Record Data Rate Can Download 'GoT' in One Second
- Digital Privacy: Europeans Threaten to Halt Data to US

"It's very dense and with the right storage conditions, DNA can be extremely long-lasting," Georg Seelig, an associate professor at the University of Washington and co-author of the study, tells *Newsweek*. "However, reading and writing DNA is still very slow, so it's good for applications where you want to keep information around for a long time but not access it often."

Ceze adds: "DNA also never becomes obsolete, unlike that old dusty floppy disk at the bottom of your drawer."

The team from the University of Washington are one of only two groups in the U.S. to have demonstrated the ability to perform "random access" to identify and retrieve data from DNA. Before this can be rolled out on a significant scale, however, the cost of synthesizing DNA—or writing it—for this purpose needs to be reduced, and the efficiency improved.

Both Ceze and Seelig agree that if the right incentives are in place, this could easily be achieved and sprawling data centers could become a thing of the past.

MOST READ


How U[...] Up As S[...] Bigges[...]


Life Ins[...] Sadisti[...] Bootca[...]


Has NA[...] Life on [...] Planets[...]


Nation[...] Carry R[...] What Is[...]


Trump [...] and 'Al[...] Preside[...]