DIGITAL DNA

COULD THE MOLECULE KNOWN FOR STORING GENETIC INFORMATION ALSO STORE THE WORLD'S DATA?

BY ANDY EXTANCE

For Nick Goldman, the idea of encoding data in DNA started out as a joke.

It was Wednesday 16 February 2011, and Goldman was at a hotel in Hamburg, Germany, talking with some of his fellow bioinformaticists about how they could afford to store the reams of genome sequences and other data the world was throwing at them. He remembers the scientists getting so frustrated by the expense and limitations of conventional computing technology that they started kidding about sci-fi alternatives. "We thought, 'What's to stop us using DNA to store information?'"

Then the laughter stopped. "It was a lightbulb moment," says Goldman, a group leader at the European Bioinformatics Institute (EBI) in Hinxton, UK. True, DNA storage would be pathetically slow compared with the microsecond timescales for reading or writing bits in a silicon memory chip. It would take hours to encode data by synthesizing DNA strings with a specific pattern of bases, and still more hours to recover that information using a sequencing machine. But with DNA, a whole human genome fits into a cell that is invisible to the naked eye. For sheer density of information storage, DNA could be orders of magnitude beyond silicon — perfect for long-term archiving.

"We sat down in the bar with napkins and biros," says Goldman, and started scribbling ideas: "What would you have to do to make that work?" The researchers' biggest worry was that DNA synthesis and sequencing made mistakes as often as 1 in every 100 nucleotides. This would render large-scale data storage hopelessly unreliable — unless they could find a workable error-correction scheme. Could they encode bits into base pairs in a way that would allow them to detect and undo the mistakes? "Within the course of an evening," says Goldman, "we knew that you could."

He and his EBI colleague Ewan Birney took the idea back to their labs, and two years later announced that they had successfully used DNA to encode five files, including Shakespeare's sonnets and a snippet of Martin Luther King's 'I have a dream' speech[1]. By then, biologist George Church and his team at Harvard University in Cambridge, Massachusetts, had unveiled an independent demonstration of DNA encoding[2]. But at 739 kilobytes (kB), the EBI files comprised the largest DNA archive ever produced — until July 2016, when researchers from Microsoft and the University of Washington claimed a leap to 200 megabytes (MB).

The latest experiment signals that interest in using DNA as a storage medium is surging far beyond genomics: the whole world is facing a data crunch. Counting everything from astronomical images and journal articles to YouTube videos, the global digital archive will hit an estimated 44 trillion gigabytes (GB) by 2020, a tenfold increase over 2013. By 2040, if everything were stored for instant access in, say, the flash memory chips used in memory sticks, the archive would consume 10–100 times the expected supply of microchip-grade silicon[3].

That is one reason why permanent archives of rarely accessed data currently rely on old-fashioned magnetic tapes. This medium packs in information much more densely than silicon can, but is much slower to read. Yet even that approach is becoming unsustainable, says David Markowitz, a computational neuroscientist at the US Intelligence Advanced Research Projects Activity (IARPA) in Washington DC. It is possible to imagine a data centre holding an exabyte (one billion gigabytes) on tape drives, he says. But such a centre would require US$1 billion over 10 years to build and maintain, as well as hundreds of megawatts of power. "Molecular data storage has the potential to reduce all of those requirements by up to three orders of magnitude," says Markowitz. If information could be packaged as densely as it is in the genes of the bacterium *Escherichia coli*, the world's storage needs could be met by about a kilogram of DNA (see 'Storage limits').

Achieving that potential won't be easy. Before DNA can become a viable competitor to conventional storage technologies, researchers

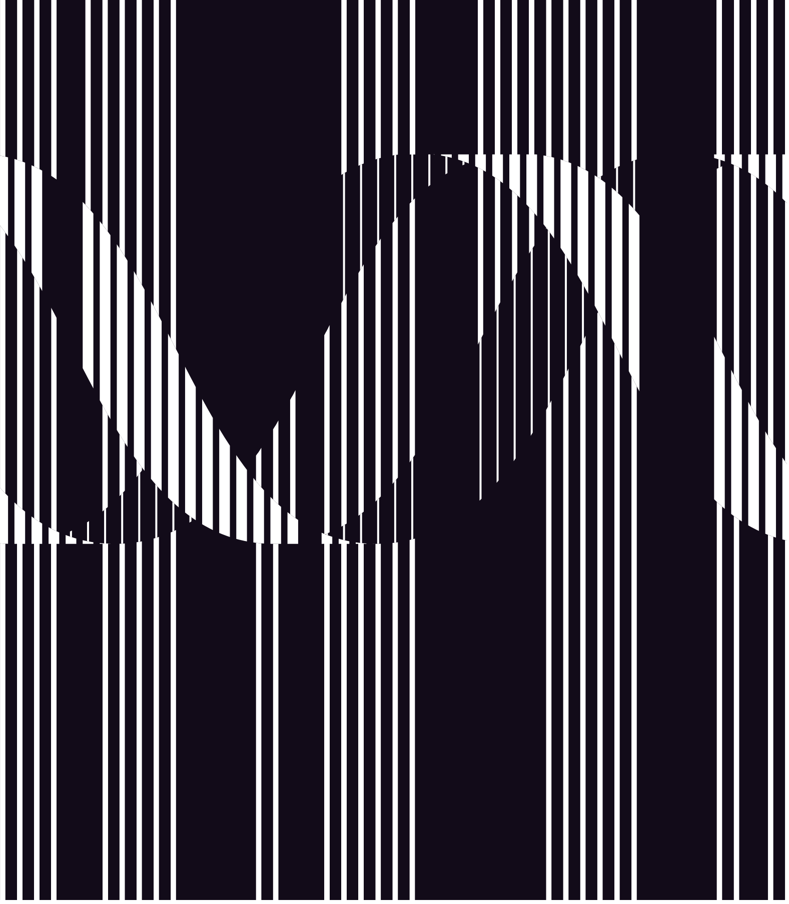


ILLUSTRATION BY WES FERNANDES/NATURE

will have to surmount a host of challenges, from reliably encoding information in DNA and retrieving only the information a user needs, to making nucleotide strings cheaply and quickly enough.

But efforts to meet those challenges are picking up. The Semiconductor Research Corporation (SRC), a foundation in Durham, North Carolina, that is supported by a consortium of chip-making firms, is backing DNA storage work. Goldman and Birney have UK government funding to experiment with next-generation approaches to DNA storage and are planning to set up a company to build on their research. And in April, IARPA and the SRC hosted a workshop for academics and industry researchers, including from companies such as IBM, to direct research in the field.

“For ten years we’ve been looking beyond silicon” for data archiving, says SRC director and chief scientist Victor Zhirnov. “It is very difficult to replace,” he says. But DNA, one of the strongest candidates yet, “looks like it may happen.”

> “ONCE WE CAN GET THOSE WRITTEN ON DNA, YOU CAN STICK IT IN A CAVE AND FORGET ABOUT IT.”

LONG-TERM MEMORY

The first person to map the ones and zeroes of digital data onto the four base pairs of DNA was artist Joe Davis, in a 1988 collaboration with researchers from Harvard. The DNA sequence, which they inserted into *E. coli*, encoded just 35 bits. When organized into a 5 × 7 matrix, with ones corresponding to dark pixels and zeroes corresponding to light pixels, they formed a picture of an ancient Germanic rune representing life and the female Earth.

Today, Davis is affiliated with Church’s lab, which began to explore DNA data storage in 2011. The Harvard team hoped the application might help to reduce the high cost of synthesizing DNA, much as genomics had reduced the cost of sequencing. Church carried out the proof-of-concept experiments in November 2011 along with Sri Kosuri, now at the University of California, Los Angeles, and genomics expert Yuan Gao at Johns Hopkins University in Baltimore, Maryland. The team used many short DNA strings to encode a 659-kB version of a book Church had co-authored. Part of each string was an address that specified how the pieces should be ordered after sequencing, with the remainder containing the data. A binary zero could be encoded by the bases adenine or cytosine, and a binary one could be represented by guanine or thymine. That flexibility helped the group to design sequences that avoided reading problems, which can occur with regions containing lots of guanine and cytosine, repeated sections, or stretches that bind to one another and make the strings fold up. They didn’t have error correction in the strict sense, instead relying on the redundancy provided by having many copies of each individual string. Consequently, after sequencing the strings, Kosuri, Church and Gao found 22 errors — far too many for reliable data storage.

At the EBI, meanwhile, Goldman, Birney and their colleagues were also using many strings of DNA to encode their 739-kB data store, which included an image, ASCII text, audio files and a PDF version of Watson and Crick’s iconic paper on DNA’s double-helix structure. To avoid repeating bases and other sources of error, the EBI-led team used a more complex scheme (see ‘Making memories’). One aspect involved encoding the data not as binary ones and zeroes, but in base three — the equivalent of zero, one and two. They then continuously rotated which DNA base represented each number, so as to avoid sequences that might cause problems during reading. By using overlapping, 100-base-long strings that progressively shifted by 25 bases, the EBI scientists also ensured that there would be four versions of each 25-base segment for error-checking and comparison against each other.

They still lost 2 of the 25-base sequences — ironically, part of the Watson and Crick file. Nevertheless, these results convinced Goldman that DNA had potential as a cheap, long-term data repository that would require little energy to store. As a measure of just how long-term, he points to the 2013 announcement of a horse genome decoded from a bone trapped in permafrost for 700,000 years[4]. “In data centres, no one trusts a hard disk after three years,” he says. “No one trusts a tape after at most ten years. Where you want a copy safe for more than that, once we can get those written on DNA, you can stick it in a cave and forget about it until you want to read it.”

A BURGEONING FIELD

That possibility has captured the imaginations of computer scientists Luis Ceze, from the University of Washington, and Karin Strauss, from Microsoft Research in Redmond, Washington, ever since they heard Goldman discuss the EBI work when they visited the United Kingdom in 2013. “DNA’s density, stability and maturity have made us excited about it,” says Strauss.

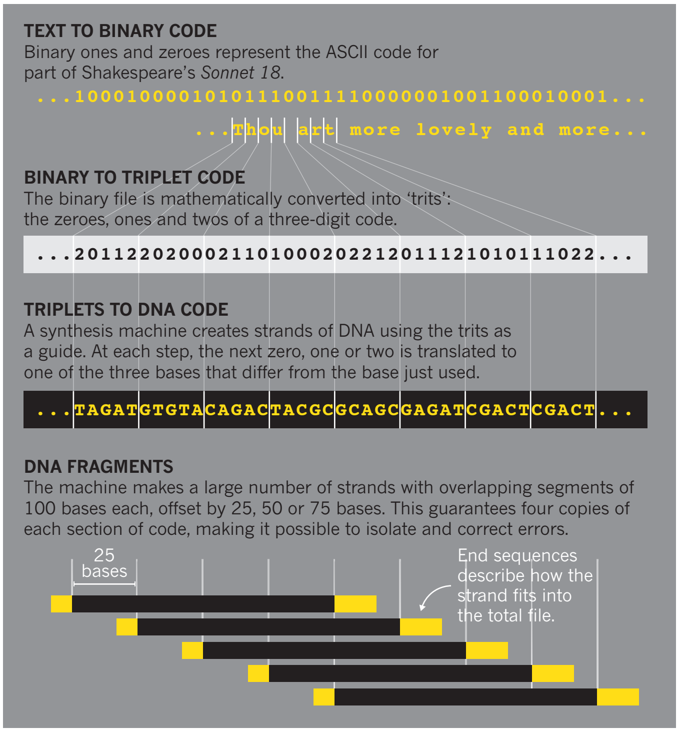
And on their return to Washington state, says Strauss, she and Ceze started investigations with their University of Washington collaborator Georg Seelig. One of their chief concerns has been another major drawback that goes well beyond DNA’s vulnerability to errors. Using standard sequencing methods, there was no way to retrieve any one piece of data without retrieving all the data: every DNA string had to be read. That would be vastly more cumbersome than conventional computer memory, which allows for random access: the ability to read just the data that a user needs.

The team outlined its solution in early April at a conference in Atlanta, Georgia. The researchers start by withdrawing tiny samples from their DNA archive. They then use the polymerase chain reaction (PCR) to pinpoint and make more copies of the strings encoding the data they want to extract[5]. The proliferation of copies makes the sequencing faster, cheaper and more accurate than previous approaches. The team has also devised an alternative error-correction scheme that the group says allows for data encoding twice as dense as the EBI’s, but just as reliable.

As a demonstration, the Microsoft–University of Washington ▶

MAKING MEMORIES

DNA DATA-ENCODING SCHEMES SUCH AS THIS ONE ARE DESIGNED TO MINIMIZE ERRORS IN SYNTHESIZING AND SEQUENCING THE MOLECULE — AND THEN CORRECT ANY ERRORS THAT DO OCCUR.



TEXT TO BINARY CODE

Binary ones and zeroes represent the ASCII code for part of Shakespeare's *Sonnet 18*.

...10001000010101110011110000001001100010001...

...Thou art more lovely and more...

BINARY TO TRIPLET CODE

The binary file is mathematically converted into 'trits': the zeroes, ones and twos of a three-digit code.

...20112202000211010002022120111210101011022...

TRIPLETS TO DNA CODE

A synthesis machine creates strands of DNA using the trits as a guide. At each step, the next zero, one or two is translated to one of the three bases that differ from the base just used.

...TAGATGTGTACAGACTACGCGCAGCGAGATCGACTCGACT...

DNA FRAGMENTS

The machine makes a large number of strands with overlapping segments of 100 bases each, offset by 25, 50 or 75 bases. This guarantees four copies of each section of code, making it possible to isolate and correct errors.

End sequences describe how the strand fits into the total file.

STORAGE LIMITS

Estimates based on bacterial genetics suggest that digital DNA could one day rival or exceed today's storage technology.

	Hard disk	Flash memory	Bacterial DNA
Read–write speed (μs per bit)	~3,000–5,000	~100	<100
Data retention (years)	>10	>10	>100
Power usage (watts per gigabyte)	~0.04	~0.01–0.04	$<10^{-10}$
Data density (bits per cm^3)	~10^{13}	~10^{16}	~10^{19}

WEIGHT OF DNA NEEDED TO STORE WORLD'S DATA

~1 kg

SOURCES: MAKING MEMORIES, REF. 1; STORAGE LIMITS, REF. 3

researchers stored 151 kB of images, some encoded using the EBI method and some using their new approach, in a single pool of strings. They extracted three — a cat, the Sydney opera house and a cartoon monkey — using the EBI-like method, getting one read error that they had to correct manually. They also read the Sydney Opera House image using their new method, without any mistakes.

ECONOMICS VERSUS CHEMISTRY

At the University of Illinois at Urbana–Champaign, computer scientist Olgica Milenkovic and her colleagues have developed a random-access approach that also enables them to rewrite the encoded data[6]. Their method stores data as long strings of DNA that have address sequences at both ends. The researchers then use these addresses to select, amplify and rewrite the strings using either PCR or the gene-editing technique CRISPR–Cas9.

The addresses have to avoid sequences that would hamper reading while also being different enough from each other to stop them being mixed up in the presence of errors. Doing this — and avoiding problems such as molecules folding up because their sequences contain stretches that recognize and bind to each other — took intense calculations. "At the beginning, we used computer search because it was really difficult to come up with something that had all these properties," Milenkovic says. Her team has now replaced this labour-intensive process with mathematical formulae that allow them to devise an encoding scheme much more quickly.

Other challenges for DNA data storage are scale and speed of synthesizing the molecules, says Kosuri, who admits that he has not been very bullish about the idea for that reason. During the early experiments at Harvard, he recalls, "we had 700 kB. Even a 1,000-fold increase on that is 700 MB, which is a CD". Truly making a difference to the worldwide data archiving problem would mean storing information by the petabyte at least. "It's not impossible," says Kosuri, "but people have to realize the scale is on the order of million-fold improvements."

That will not be easy, agrees Markowitz. "The dominant production method is an almost 30-year-old chemical process that takes upwards of 400 seconds to add each base," he says. If this were to remain the approach used, he adds, billions of different strings would have to be made in parallel for writing to be fast enough. The current maximum for simultaneous production is tens of thousands of strings.

A closely related factor is the cost of synthesizing DNA. It accounted for 98% of the expense of the $12,660 EBI experiment. Sequencing accounted for only 2%, thanks to a two-millionfold cost reduction since the completion of the Human Genome Project in 2003. Despite this precedent, Kosuri isn't convinced that economics can drive the same kind of progress in DNA synthesis. "You can easily imagine markets to sequence 7 billion people, but there's no case for building 7 billion people's genomes," he says. He concedes that some improvement in costs might result from Human Genome Project-Write (HGP-write), a project proposed in June by Church and others. If funded, the programme would aim to synthesize an entire human genome: 23 chromosome pairs containing 3.2 billion nucleotides. But even if HGP-write succeeds, says Kosuri, a human genome contains just 0.75 GB of information and would be dwarfed by the challenge of synthesizing practical data stores.

Zhirnov, however, is optimistic that the cost of synthesis can be orders of magnitude below today's levels. "There are no fundamental reasons why it's high," he says.

In April, Microsoft Research made an early move that may help create the necessary demand, ordering 10 million strings from Twist Bioscience, a DNA synthesis start-up company in San Francisco, California. Strauss and her colleagues say they have been using the strings to push their random-access storage approach to 0.2 GB. The details remain unpublished, but the archive reportedly includes the Universal Declaration of Human Rights in more than 100 languages, the top 100 books of Project Guttenberg and a seed database. Although this is much less of a synthesis challenge than the HGP-write faces, Strauss stresses the significance of the 250-fold jump in storage capacity.

"It was time to exercise our muscle handling larger volumes of DNA to push it to a larger scale and see where the process breaks," she says. "It actually breaks in multiple places — and we're learning a great deal out of it."

Goldman is confident that this is just a taste of things to come. "Our estimate is that we need 100,000-fold improvements to make the tech-nology sing, and we think that's very credible," he says. "While past performance is no guarantee, there are new reading technologies com-ing onstream every year or two. Six orders of magnitude is no big deal in genomics. You just wait a bit." ■

Andy Extance *is a freelance writer in Exeter, UK.*

1. Goldman, N. *et al. Nature* **494,** 77–80 (2013).
2. Church, G. M., Gao, Y. & Kosuri, S. *Science* **337,** 1628 (2012).
3. Zhirnov, V., Zadegan, R. M., Sandhu, G. S., Church, G. M. & Hughes, W. L. *Nature Mater.* **15,** 366–370 (2016).
4. Orlando, L. *et al. Nature* **499,** 74–78 (2013).
5. Bornholt, J. *et al.* in *Proc. 21st Int. Conf. Archit. Support Program. Lang. Oper. Syst.* **44,** 637–649 (ACM, 2016).
6. Hossein Tabatabaei Yazdi, S. M., Yuan, Y., Ma, J., Zhao, H. & Milenkovic, O. *Sci. Rep.* **5,** 14138 (2015).

CORRECTION

The News Feature ‘Digital DNA’ (*Nature* **537,** 22–24; 2016) gave the an incorrect size for the 2013 EBI files. The correct figure is 739 kilobytes not 739 kilobases.