Storing information in DNA

Test-tube data

Comment (2) | Print

E-mail | Reprints & permissions

Archives could last for thousands of years when stored in DNA instead of magnetic tapes and hard drives

Jan 26th 2013 | From the print edition

LIKE all the best ideas, this one was born in a pub. Nick Goldman and Ewan Birney of the European Bioinformatics Institute (EBI) near Cambridge, were pondering what they could do with the torrent of genomic data their research group generates, all of which has to be archived.

The volume of data is growing faster than the capacity of the hard drives used to hold it. "That means the cost of storage is rising, but our budgets are not," says Dr Goldman. Over a few beers, the pair began wondering if artificially constructed DNA might be one way to store the data torrent generated by the natural stuff. After a few more drinks and much scribbling on beer mats, what started out as a bit of amusing speculation had turned into the bones of a workable scheme. After some fleshing out and a successful test run, the full details were published this week in *Nature*.

The idea is not new. DNA is, after all, already used to store information in the form of genomes by every living organism on Earth. Its prowess at that job is the reason that information scientists have been trying to co-opt it for their own uses. But this has not been without problems.

Related topics

- Technology
- Computer hardware
- Data storage
- Science and technology
- Computer technology

Dr Goldman's new scheme is significant in several ways. He and his team have managed to set a record (739.3 kilobytes) for the amount of unique information encoded. But it has been designed to do far more than that. It should, think the researchers, be easily capable of swallowing the roughly 3 zettabytes (a zettabyte is one billion trillion or 10^{21} bytes) of digital data thought presently to exist in the world and still have room for plenty more. It would do so with a density of around 2.2 petabytes (10^{15}) per gram; enough, in other words, to fit all the world's digital information into the back of a lorry. Moreover, their method dramatically reduces the copying errors to which many previous DNA storage attempts have been prone.

Latest blog posts - All times are GMT

Barack Obama's second-term strategy: The long game
Democracy in America - 23 mins ago

North Korean propaganda: Human pixels
Prospero - 3 hours 5 mins ago

North Korean sanctions: Nuclear reaction
Banyan - Jan 24th, 08:28

Recommended economics writing: Link exchange
Free exchange - Jan 23rd, 22:29

Justice in Mexico: Florence goes free
Americas view - Jan 23rd, 22:26

Criminal justice and the courts: Thumb on the scale
Democracy in America - Jan 23rd, 21:58

Syria's war: The axis power
Pomegranate - Jan 23rd, 20:49

More from our blogs »

Most popular

Recommended | Commented

Faithful reproduction

The trick to this fidelity lies in the way the researchers translate their files from the hard drive to the test tube. DNA uses four chemical "bases"—adenosine (A), thymine (T), cytosine (C) and guanine (G)—to encode information. Previous approaches have often mapped the binary 1s and 0s used by computers directly onto these bases. For instance, A and C might represent 0, while G and T signify 1. The problem is that sequences of 1s or 0s in the source code can generate repetition of a single base in the DNA (say, TTTT). Such repetitions are more likely to be misread by DNA-sequencing machines, leading to errors when reading the information back.

The team's solution was to translate the binary computer information into ternary (a system that uses three numerals: 0, 1 and 2) and then encode that information into the DNA. Instead of a direct link between a given number and a particular base, the encoding scheme depends on which base has been used most recently (see table). For instance, if the previous base was A, then a 2 would be represented by T. But if the previous base was G, then 2 would be represented by C. Similar substitution rules cover every possible combination of letters and numbers, ensuring that a sequence of identical digits in the data is not represented by a sequence of identical bases in the DNA, helping to avoid mistakes.

No repetition, please

A DNA coding scheme

Previous base written	Digit to be encoded		
	0	*1*	*2*
A	C	G	T
C	G	T	A
G	T	A	C
T	A	C	G

Source: Nick Goldman *et al*, *Nature*

The code then had to be created in artificial DNA. The simplest approach would be to synthesise one long DNA string for every file to be stored. But DNA-synthesis machines are not yet able to do that reliably. So the researchers decided to chop their files into thousands of individual chunks, each 117 bases long. In each chunk, 100 bases are devoted to the file data themselves, and the remainder used for indexing information that records where in the completed file a specific chunk belongs. The process also contains the DNA equivalent of the error-detecting "parity bit" found in most computer systems.

To provide yet more tolerance for mistakes, the researchers chopped up the source files a further three times, each in a slightly different, overlapping way. The idea is to ensure that each 25-base quarter of a 100-base chunk was also represented in three other chunks of DNA. If any copying errors did occur in a particular chunk, it could be compared against its three counterparts, and a majority vote used to decide which was correct. Reading the chunks back is simply a matter of generating multiple copies of the fragments using a standard chemical reaction, feeding these into a DNA-sequencing machine and stitching the files back together.

When the scheme was tested, it worked almost as planned. The researchers were able to encode and decode five computer files, including an MP3 recording of part of Martin Luther King's "I have a dream" speech and a PDF version of the 1953 paper by Francis Crick and James Watson describing the structure of DNA. The one glitch was that, despite all the precautions, two 25-base segments of the DNA paper went missing. The problem was eventually traced to a combination of a quirk of DNA chemistry and another quirk in the machines used to do the synthesis. Dr Goldman is confident that a tweak to their code will avoid the problem in future.

There are downsides to DNA as a data-storage medium. One is the relatively slow speed at which data can be read back. It took the researchers two weeks to reconstruct their five files, although with better equipment it could be done in a day. Beyond that, the process can be sped up by adding more sequencing machines.

Ironically, then, the method is not suitable for the EBI's need to serve up its genome data over the internet at a moment's notice. But for less intensively used archives, that might not be a problem. One example given is that of CERN, Europe's biggest particle-physics lab, which maintains a big archive of data from the Large Hadron Collider.

Store out of direct sunlight

1 **The Senkaku/Diaoyu islands**
Dangerous shoals

2 **Aaron Swartz**
3 **Barack Obama**: How will history see me?
4 **Revamping Skopje**: Stones of contention
5 **Politics this week**

The other disadvantage is cost. Dr Goldman estimates that, at commercial rates, their method costs around $12,400 per megabyte stored. That is millions of times more than the cost of writing the same data to the magnetic tape currently used to archive digital information. But magnetic tapes degrade and must be replaced every few years, whereas DNA remains readable for tens of thousands of years so long as it is kept somewhere cool, dark and dry—as proved by the recovery of DNA from woolly mammoths and Neanderthals.

The longer you want to store information, then, the more attractive DNA becomes. And the cost of sequencing and synthesising DNA is falling fast. The researchers reckon that, within a decade, that could make DNA competitive with other methods for (infrequently-used) archives designed to last fifty years or more.

There is one final advantage in using DNA. Modern, digital storage technologies tend to come and go: just think of the fate of the laser disc, for example. In the early 2000s NASA, America's space agency, was reduced to trawling around internet auction sites in order to find old-style eight-inch floppy drives to get at the data it had laid down in the 1960s and 1970s. But, says Dr Goldman, DNA has endured for more than 3 billion years. So long as life—and biologists—endure, someone should know how to read it.

From the print edition: Science and technology

Related items

TOPIC: Technology »

Babbage: January 23rd 2013: Watch out Google

Telecommunications: Nearing the end of the line

High-definition TV: Difference Engine: Ne plus ultra

TOPIC: Computer hardware »

Electronic devices on planes: Turn off your iPad now, please

Tablet computers: Difference Engine: Smaller still is smarter

Supercomputing: Deeper thought

TOPIC: Data storage »

Phase-change memory: Altered states

Printing in DNA: Words in a vial

Privacy: Mad crush

TOPIC: Science and technology »

Flu research: Back in business

A brief history of macro: How we got here

China's workforce: Will you still need me?

More related topics: Computer technology