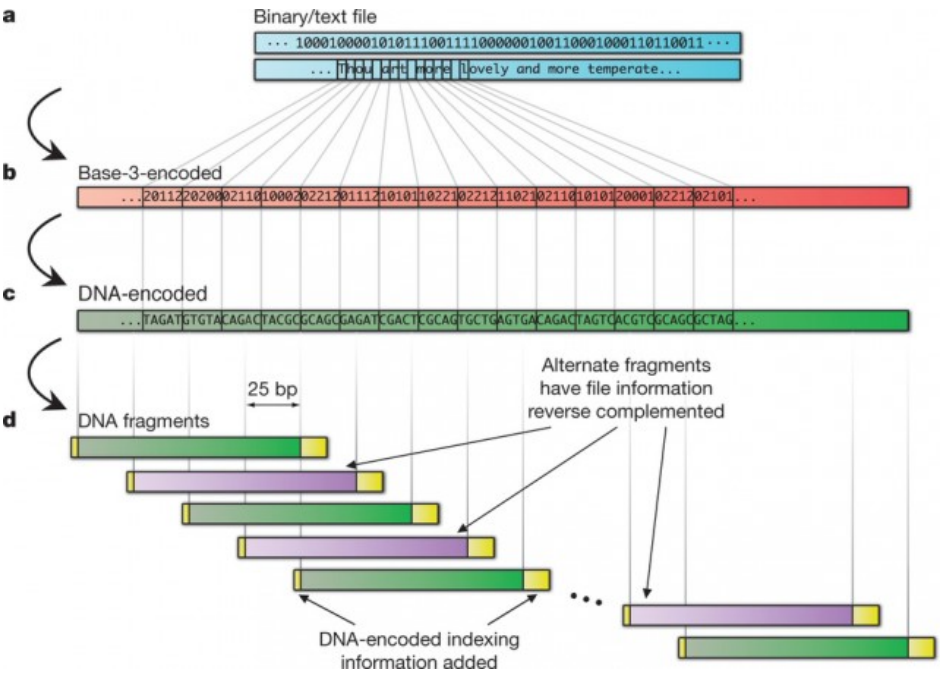
MP3 files written as DNA with storage density of 2.2 petabytes per gram

Researchers used trinary to take advantage of DNA's four bases.

by **John Timmer** - Jan 23 2013, 6:20pm GMT

LIFE SCIENCES 48



The general approach to storing a binary file as DNA, described in detail below.

Goldman et al., Nature

It's easy to get excited about the idea of encoding information in single molecules, which seems to be the ultimate end of the miniaturization that has been driving the electronics industry. But it's also easy to forget that we've been beaten there—by a few billion years. The chemical information present in biomolecules was critical to the origin of life and probably dates back to whatever interesting chemical reactions preceded it.

It's only within the past few decades, however, that humans have learned to speak DNA. Even then, it took a while to develop the technology needed to synthesize and determine the sequence of large populations of molecules. But we're there now, and people have started experimenting with putting binary data in biological form. Now, a new study has confirmed the flexibility of the approach by encoding everything from an MP3 to the decoding algorithm into fragments of DNA. The cost analysis done by the authors suggest that the technology may soon be suitable for decade-scale storage, provided current trends continue.

Trinary encoding

Computer data is in binary, while each location in a DNA molecule can hold any one of four bases (A, T, C, and G). Rather than using all that extra information capacity, however, the authors used it to avoid a technical problem. Stretches of a single type of base (say, TTTTT) are often not sequenced properly by current techniques—in fact, this was the biggest source of errors in the previous DNA data storage effort. So for this new encoding, they used one of the bases to break up long runs of any of the other three.

(To explain how this works practically, let's say the A, T, and C encoded information, while G represents "more of the same." If you had a run of four A's, you could represent it as AAGA. But since the G doesn't encode for anything in particular, TTGT can be used to represent four T's. The only thing that matters is that there are no more than two identical bases in a row.)

TOP FEATURE STORY

FEATURE STORY (2 PAGES)

Review: Acer's Iconia W700 is an Ultrabook in a tablet's body

The tablet balances size, performance, and battery life with some success.

71

STAY IN THE KNOW WITH

LATEST NEWS

New features, old franchises coming to the Wii U

MISSED IT BY THAT MUCH

If Verizon had its way, Siri would have been Android's marquee feature

THE PIXELS, DUKE

Nokia's next-gen Windows Phone may get a 41-megapixel camera

PROPOSITUM PECUNIAE EST EXPENDENDO

The Dealmaster: Twenty-seven inches of IPS love

PLEASE DON'T LEAVE US

RIM tries to keep its business customers ahead of BlackBerry 10 launch

That leaves three bases to encode information, so the authors converted their information into trinary. In all, they encoded a large number of works: all 154 Shakespeare sonnets, a PDF of a scientific paper, a photograph of the lab some of them work in, and an MP3 of part of Martin Luther King's "I have a dream" speech. For good measure, they also threw in the algorithm they use for converting binary data into trinary.

Once in trinary, the results were encoded into the error-avoiding DNA code described above. The resulting sequence was then broken into chunks that were easy to synthesize. Each chunk came with parity information (for error correction), a short file ID, and some data that indicates the offset within the file (so, for example, that the sequence holds digits 500-600). To provide an added level of data security, 100-bases-long DNA inserts were staggered by 25 bases so that consecutive fragments had a 75-base overlap. Thus, many sections of the file were carried by four different DNA molecules.

And it all worked brilliantly—mostly. For most of the files, the authors' sequencing and analysis protocol could reconstruct an error-free version of the file without any intervention. One, however, ended up with two 25-base-long gaps, presumably resulting from a particular sequence that is very difficult to synthesize. Based on parity and other data, they were able to reconstruct the contents of the gaps, but understanding why things went wrong in the first place would be critical to understanding how well suited this method is to long-term archiving of data.

Long-term storage

In general, though, the DNA was very robust. The authors simply dried it out before shipping it to a lab in Germany (with a layover in the UK), where it was decoded. Careful storage in a cold, dry location could keep it viable for much, much longer. The authors estimate their storage density was about 2.2 Petabytes per gram, and that it included enough DNA to recover the data about ten additional times.

Which brings us to the authors' more general argument. Assuming the process is streamlined and automated, and the physical cataloging can be handled at minimal cost, is this ever likely to be a cost-effective way to store data? As a point of contrast, the authors considered a data set from the LHC. After a few years, access to data from these archives tends to be very limited, while the cost of maintaining them tends to involve sporadic migrations to upgraded magnetic tape technology.

With current, state-of-the-art DNA synthesis and sequencing, the economics start to make sense only if you're planning on storing the data for over 500 years (although it would be good out to 5,000). But they also note that if the relevant technologies continue improving at their current rates, it will only take a decade to get to where DNA-based storage would start making sense for archiving data for as little as 50 years.

Both this team and the previous one point out that, unlike storage media, the actual physical storage of information won't change if DNA is used, even if we end up using different methods to synthesize or read out the molecule's contents. And life on Earth will always ensure that whatever we need to manipulate it will always be available to us. In other words, if we're ever *not* able to read the information in DNA, then we've got bigger problems than having lost some data.

Nature, 2013. DOI: 10.1038/nature11875 (About DOIs).

John Timmer / John became Ars Technica's science editor in 2007 after spending 15 years doing biology research at places like Berkeley and Cornell.

Follow @j_timmer

← OLDER STORY

YOU MAY ALSO LIKE