Storing data as DNA – as easy as ACTG?

FEATURE | JANUARY 15, 2014 | BY CHARLES HARVEY

iSGTW recently interviewed Ewan Birney , associate director of the European Bioinformatics Institute (EBI) , regarding his keynote talk at the EUDAT 2nd Conference . In this interview, Birney raised the exciting prospect of using DNA as an organic data storage device. But could DNA storage really replace tapes and hard disks for long-term preservation of data? Charles Harvey investigates…

"DNA is remarkable: just one gram of DNA can store about a petabyte's worth of data, and that's with the redundancy required to ensure that it's fully error tolerant. It's estimated that you could put the whole internet into the size of a van! You can also copy trivially. The only problem at the moment is cost: it's prohibitively expensive to write DNA. Nevertheless, this technology is expected to come down in price dramatically over the coming years. The only question is: how quickly will it come down in price?

DNA could be really useful for storing data sets on a really long timescale. The earliest writings ever discovered are under 6,000 years old, so data stored on DNA would outlast everything we know about today. DNA data storage could be really useful for storing things like films, governmental archives, key scientific data, and so on. For example, it could be really useful for climate change research to store an archive of Earth-observing satellite images from previous decades."

***Read more:** Data in the DNA: transforming biology and data storage.*

Image courtesy Duncan Hull, Flickr (CC BY-SA 2.0).

Front page image courtesy Micah Baldwin, Flickr (CC BY-SA 2.0).

DNA is the world's oldest data storage device. The technology to read and write DNA has become commonplace since bacteria were first genetically engineered in 1973. And, while it's possible to store petabytes of data in a microscopic space, might it ever be worthwhile to store information as DNA, rather than on hard drives or magnetic tape?

To store information, DNA uses 4 bases (Adenine, Cytosine, Thymine, and Guanine — often simply referred to as A, C, T, and G). In 2003, Pak Chung Wong from the Pacific Northwest National Laboratory, USA, encrypted text into DNA by converting each character into a base-4 sequence of numbers, each corresponding to a certain base. Using genetic engineering, these sequences were inserted into the genome of a bacterium, once repetitive sequences of numbers — space-consuming and potentially harmful for the bacteria — had been removed. Special beginning and end tags were also added to the strands to allow indexing and to prevent the bacteria expunging the inserted sequences as viral invaders.

Bacteria are an obvious option when it comes to storing data as DNA: they replicate quickly, creating numerous copies of the data in the process. Also, should a mutation occur within an individual bacterium, the remaining bacteria will still contain the original information, allowing researchers to recover the original sequence with close-to-perfect accuracy.

Deinococcus radiodurans , a bacterial species adapted to survive in extreme conditions, was chosen by Wong and his colleagues to be the host due to its ability to quickly repair spontaneously arising mutations. Unlike hard drives or magnetic tapes, which are vulnerable to physical damage, data stored in bacteria could survive numerous natural disasters and be safely passed on to future generations.

However, the heterologous (artificially inserted) DNA could make the bacterial genome unstable, believes Geoff Baldwin, reader in biochemistry at Imperial College London, UK. "Bacteria are highly evolved organisms with relatively minimal genomes", says Baldwin. "There is always the issue that maintaining large quantities of heterologous DNA will exert a fitness burden that will favor loss of the additional DNA, which does not bode well for the use of bacteria as a mass data storage device."

While mutation rates are relatively low (approximately 1 base every 10,000 generations), bacteria's fast replication rate could make long-term data storage problematic. Another issue is that if the inserted DNA is similar to that of the host bacteria, it could interfere with its normal cellular processes. "Ultimately this means that there is not complete freedom to insert any sequence," says Baldwin. "This can be overcome to some extent depending on the method chosen to encode information, but there always remains the possibility for instability due to unexpected consequences."

While data storage in bacteria isn't yet sufficiently developed to be used for mass storage, using 'naked' DNA could be a more promising alternative. Mammoths and Neanderthals have been found preserved for thousands of years with DNA sequences intact, showing that living cells are not required for DNA itself to remain an efficient, stable data-storage means. Naked DNA is easier to use than bacteria as it doesn't require genetic manipulations to safely insert it into a host.

Birney was part of a team that encoded a record-breaking 700 kilobytes of unique data — including all 154 of Shakespeare's sonnets — into naked DNA and retrieved it with 99% accuracy. Translating binary data into a 4-base system often results in long sequences of identical bases, which have a tendency to be misread by DNA-sequencing machines, ultimately degrading the information of the original message. The team came up with an ingenious system that allowed them to encode data with such high fidelity. They used a base-3 encoding system: depending on which base was last encoded, a 0, 1 or 2 would correspond with one of the 3 other bases, ensuring the creation of a sequence void of any repetition.

Despite the current high cost of writing and reading DNA, Christophe Dessimoz, another member of the research team, remains hopeful as to the future use of data storage in DNA. "Our analysis shows that for small quantities of data, it is already economically viable for very long term (1,000 years or more), which is relevant for applications such as storing the location of nuclear sites," says Dessimoz, who is now based at University College London, UK. "If the current pace of technological development continues, within the next decade DNA-based storage will become economical for applications with time horizons of 50 years or more."

Documents that need to be kept long-term but which are seldom used, such as governmental, legal and scientific archives, could make use of DNA storage today. Currently, these archives require transferring to newer devices every few years — at great expense — in order to preserve them. Altering the information without discarding and rewriting the DNA, however, isn't yet possible. And although for now, the entire DNA strand needs reading to find a specific part, this might someday become feasible. "Our study was not concerned with the problem of reading subsets of data," explains Dessimoz, "but as we can store arbitrary information, it would indeed be possible to store an index and to partition the data into separate (presumably very small) physical containers, *e.g.* on a chip."

Average:

Your rating: None Average: 3 (1 vote)

About the Author »

Charles Harvey

Charles Harvey is a freelance science writer.

RELATED TERMS: amino acids data DNA EBI EUDAT Europe genetics organic Shakespeare storage biology computer science information science

Comments

ADD NEW COMMENT

Post new comment

Subject:

Comment: *

By submitting this form, you accept the Mollom privacy policy.

SAVE PREVIEW