Data storage in synthetic DNA

Bryan Dyne, Shane Feratu
25 July 2013

New research, published in the journal *Nature* by Nick Goldman and his colleagues, has combined genetics and computer science to develop a technique that allows any type of data to be reliably stored in the molecule deoxyribonucleic acid (DNA), offering a way to archive data over spans of decades, centuries and millennia that previously had never been practical or cost-efficient. The data that was stored totaled 739 kilobytes and the original files were retrieved with 100 percent accuracy.

As proof of concept, Goldman's team stored all of Shakespeare's 154 sonnets (text format) and another text file; the paper by James Watson and Francis Crick that first described the double-helical nature of DNA (pdf format); an audio excerpt from Martin Luther King's 1963 "I have a dream" speech (MP3 format); and a color photograph (jpeg format).

The first bits of digital information were stored in DNA in 1988. However, it was impractical, both difficult and expensive. Even as recently as 2012, costs and times for reading and writing were only practical for century-scale archives. The latest development makes storing data on a 50-year timescale feasible, and archives for shorter timescales could become cost effective within a decade.

Archiving digital data is one of the computer world's great problems. Currently, the amount of data in the world is doubling every two years. By 2020, it is estimated that this will amount to 40 trillion gigabytes of digital information. Alongside this, however, the amount of data that can be stored in a given space has also increased. In 2007, the highest information density surpassed 1 trillion bits of information per square inch. Storage is a problem, but a manageable one.

A much more difficult problem is the constant need to maintain digital systems. Degradation occurs both as a natural consequence of storing information digitally (i.e., hard-disks eventually just stop working) and the constant evolution of digital media. Important data could be stored on old floppy disks, for example, but since modern computers, especially laptops, don't have a way of reading them, the data stored is useless. In addition, when data is transferred from one device to another, there is always some data loss which is noticeable not over years, but over decades and centuries.

DNA has been sought after as a way to solve both the main problem of maintaining digital systems as well as the need to store ever-greater amounts of data. Entire genomes have been sequenced even after the degradation of tens of thousands or even hundreds of thousands of years. At its theoretical maximum capacity, a single gram of single-stranded DNA can encode 455 billion gigabytes of data. Most importantly, as shown by its essential role in biology, DNA already has the chemical composition to read and write information repeatedly and still be clearly recoverable.

Information is stored on DNA by encoding bits onto nucleotides. Nucleotides are macromolecules that are the building blocks of DNA and its companion molecule ribonucleic acid (RNA). DNA has four different types of bases, split into single-ringed (pyrimidines) and double-ringed (purines) structures: adenine, guanine and cytosine, thymine (or uracil in RNA). To actually make DNA, the four nucleotides bond together, alanine with thymine and guanine with cytosine, to form a double helix.

Naturally occurring, DNA codes the identity of a cell and passes it on from one generation to the next. It controls the making of itself and other proteins, making even minor mutations or changes in itself significant in the expression of the characteristics of the cell and on a larger scale in a species.

To get the information from the DNA strand, a process called transcription has to occur. For that the DNA helix is separated by the molecule RNA

polymerase, which then binds to part of the DNA and makes a template. This forms a complementary DNA strand, also known as mRNA or messenger RNA. After the mRNA is processed a bit further, it is ready to use what it copied from the DNA in whatever way the cell needs. Any errors in replication, such as placing the wrong nucleotide or deleting one, are generally repaired by automatic biochemical processes.

Goldman and his team used a somewhat different process to encode digital data onto DNA, but the fundamentals remain the same. Data is written onto nucleotides, stored and then retrieved when needed. What is significant about the latest developments is that any sort of data can be stored, whereas before only specific applications were realized.

At the most elementary level, every file on a computer is a sequence of 0s and 1s. However, this is determined more from how computers themselves work and is not a strict imitation on how data can be stored. As such, the sequence of 0s and 1s can be translated into a sequence of 0s, 1s and 2s, much like how English can be translated in Spanish. What is said is the same but how it is said is different.

The coded data is then translated into the bases of genetic material adenine, guanine and cytosine, thymine (A, G, C, T). For example, the series of 1s and 0s that represent the letter “T” becomes the DNA sequence TAGAT. Words then become strands of five DNA letters together. “Thou” from the beginning of Shakespeare’s sonnet 18 becomes TAGATGTGTACAGACTACGC. The process can be mimicked for MP3s, pdfs or any other file type that exists.

One of the other benefits of this latest technique is built-in redundancy. Each “word” can be written into different DNA strands four different ways, meaning that the possibility of losing any data is extremely remote. Losing data during synthetic transcription was one of the reasons that earlier techniques for storing data on DNA were infeasible. This obstacle has now been overcome.

The last problem that has yet to be overcome is the speed at which data in DNA is accessed. So far, the speed of reading and writing is not comparable to current digital technology, but every effort is being made to shorten read and write times as much as possible not just for DNA data storage, but also to sequence genomes generally.

The great potential to archive large quantities of data in DNA is a new opportunity in man’s understanding of nature. For the first time since mankind has been recording discoveries about the natural world, we have the capability to preserve the data easily for generations, ensuring all the current culture of humanity can be appreciated by future humans.