

Shakespeare's sonnets encoded in DNA

All 154 Shakespeare sonnets were spelled out in DNA to demonstrate the vast potential of genetic data storage

Ian Sample, science correspondent

The Guardian, Wednesday 23 January 2013 19.21 GMT



When written in DNA, one of Shakespeare's sonnets weighs 0.3 millionths of a millionth of a gram. Photograph: Oli Scarff/Getty

His words have touched the lovelorn and been pored over by brooding teenagers for more than four hundred years, but now some of the most romantic poems ever penned have been written into the code of life.

The entire collection of Shakespeare's 154 sonnets has been spelled out in DNA by scientists in Cambridge to demonstrate the vast potential of genetic storage. Huge quantities of information could be written into specks of DNA and archived for tens of thousands of years, the researchers claim.

Alongside the Bard's sonnets, the scientists made strands of DNA that stored part of an audio file of [Martin Luther King's](#) 1963 speech "I have a dream", and the [seminal research paper](#) that first described the double helical nature of DNA by Francis Crick and James Watson, a decade earlier.

Written in DNA, one of Shakespeare's sonnets weighs 0.3 millionths of a millionth of a

gram. One gram of DNA could hold as much information as more than a million CDs, the researchers said.

Nick Goldman and Ewan Birney at the European Bioinformatics Institute in Hinxton, near Cambridge, came up with the idea in a pub in Hamburg. They wondered what alternatives might exist to the expensive hard disks and magnetic tapes used to store the growing datasets that are becoming ever more common in biology.

They knew that DNA was an incredibly efficient and compact way to store information, and set about devising a way to turn the molecules into digital memory: capable of encoding the 1s and 0s used to store words, images, music and video on computers.

"We wrote on napkins and sketched out details, and realised we could probably do this," said Goldman.

The scientists developed a code that used the four molecular letters or "bases" of genetic material – known as G, T, C and A – to store information.

Digital files store data as strings of 1s and 0s. The Cambridge team's code turns every block of eight numbers in a digital code into five letters of DNA. For example, the eight digit binary code for the letter "T" becomes TAGAT.

To store words, the scientists simply run the strands of five DNA letters together. So the first word in "Thou art more lovely and more temperate" from Shakespeare's sonnet 18, becomes TAGATGTGTACAGACTACGC.

To test the procedure, Goldman and Birney converted the sonnets, Martin Luther King's speech, and the Watson and Crick paper into their new DNA code and emailed it to a US company called Agilent, which makes physical strands of DNA commercially for researchers. Several weeks later, the company returned a tiny test tube to the scientists. It contained a speck of DNA that encoded all the information they had sent.

To make sure the DNA stored the information correctly, the Cambridge scientists mixed it into a solution and ran it through a gene sequencing machine. From that, they were able to read the complete files again, according to a report in Nature.

A similar feat was achieved by the Harvard geneticist, George Church, last year. He encoded a copy of his book *Regenesis*, 11 images, and a computer program in DNA.

The Cambridge work differs in how it corrects for spelling mistakes made when the DNA is either written, or read out at the other end. Each "word" of DNA code is written in four different ways, and there are millions of copies of each of these. Any mistakes in the DNA are likely to affect only a minority of the strands of code, and these can be spotted by comparing them.

"One of the great properties of DNA is that you don't need any energy to store it. You keep it cold, dry and dark, and it lasts for a very long time. We know that because we routinely sequence woolly mammoth DNA that is kept by chance in those sorts of conditions," said Birney.

The high cost of making DNA today means genetic storage in biological hard drives will not be practical for a long time. But Birney said that another 100-fold drop in the cost of DNA sequencing, which has already happened in the past decade, makes DNA storage economic for information that has to be archived for more than 50 years.

Asked whether the DNA might pose a danger to health, if it found its way into a person or other animal, Goldman said: "The DNA we've created can't be incorporated accidentally into a genome, it uses a completely different code to that used by the cells of living bodies. If you did end up with any of this DNA inside you it would just be degraded and disposed of."
