Long-term data storage in DNA

Jonathan P.L. Cox

This article discusses how DNA might be used to store data. It is argued that, at present, DNA would be best employed as a long-term repository (thousands or millions of years). How data-containing DNA might be packaged and how the data might be encrypted, with particular attention to the encryption of written information, is also discussed. Various encryption issues are touched on, such as how data-containing DNA might be differentiated from genetic material, error detection, data compression and reading frame location. Finally, this article broaches the difficulty of constructing very large pieces of DNA in the laboratory and highlights some complications that might arise when attempting to transmit DNA-encrypted data to recipients who are a long period of time in the future.

Over the past 30 years it has been suggested several times, with varying degrees of seriousness, that DNA might be used to store data[1–3] (see also http://www.nytimes.com/library/magazine/millennium). As a storage material, DNA has some excellent qualities: (1) it has a proven track record in this area (life); (2) suitably protected (in a spore, for instance) it can be preserved for some time, probably millions of years; (3) it is self-reproducing; and (4) a substantial amount of information can be placed in its base sequence. In fact, character for character, the storage density (characters m^{-2}) of a spore of the bacterium *Bacillus subtilis* (genome size 4.2 Mbp, diameter 1 μm) is, approximating it to a flat disk, just over 20 million times that of a 250 Mbyte ZIP disk (diameter 10 cm). It is also worth pointing out that for sequences and strings of the same length (i.e. comprising the same number of characters), the DNA sequence is inherently more information-rich than the string of binary digits (there are 2^N times as many DNA sequences of length N nucleotides as there are strings N binary digits long).

One major disadvantage of DNA as a storage medium, however, is the time it would take to retrieve the data from the nucleotide sequence. Even if it was possible to feed DNA, without any preparation, into a hypothetical sequencer that read the sequence at enzymatic rates (say 100 nucleotides s^{-1}), the retrieval process would still be six orders of magnitude slower than that of a personal computer (which can read data from the hard drive at ~100 Mbits s^{-1}). Consequently, DNA is unlikely to compete with optical or magnetic (or quantum) formats in the foreseeable future. The fact that it is also unlikely ever to become obsolete[3], however, would make it ideal for the really long-term deposition of information, the sort that a human civilisation thousands or millions of years in the future might find useful. Almost certainly there will be some debate about what information a future civilisation might find useful, but surely meteorological, zoological, geological and astrophysical data would be strong candidates for inclusion. For example, meteorological data might assist the prediction of climate change and similarly descriptions of existing species, including their genomes[4,5], might help to document the course of evolution (or mass extinction). Most current palaeontologists would have been grateful for a definitive guide to the origin of birds, had it been possible to leave one at the time (although having this information would make life less interesting for onlookers of the controversies that such issues generate[6]).

Jonathan P.L. Cox
Dept of Chemistry,
University of Bath,
Bath, UK BA2 7AY.
e-mail: j.p.l.cox@bath.ac.uk

Packaging DNA for long-term storage

As aforementioned, the best format for storing DNA is probably in a spore. Spores can withstand extremely hostile conditions[7] (as might be encountered during the long storage period) and the recent painstaking extractions of bacterial species from amber and salt[8,9] suggest that they could be revived after millions of years. Furthermore, on reviving the spores by culturing them, millions of copies of the message contained within each spore would be produced.

'Most current paleontologists would have been grateful for a definitive guide to the origin of birds...'

Two spore-forming microorganisms that would be suitable hosts for the message are *Bacillus subtilis* and bakers' yeast, *Saccharomyces cerevisiae*. The molecular genetics of both species is well established. However, although much is known about the spore resistance of *B. subtilis*[7], hardly anything is known about that of *S. cerevisiae*. Conversely, *S. cerevisiae* can accommodate ten times more foreign DNA than can *B. subtilis*. To compensate for the disadvantages of each organism, the best strategy would be to reproduce the message in both microorganisms, while at the same time thoroughly investigating the spore resistance of *S. cerevisiae*.

Additional protection could be given to the spore-encased DNA by implanting the spores in amber capsules or in an artificial resin of similar composition. Samples of the microorganisms without the message should also be included as 'internal standards', in case their free living forms change radically during the storage period, preventing identification of the foreign, message DNA.

Encrypting the message

Although DNA could also be used to encrypt pictorial and numerical data, we will concentrate on how it

0167-7799/01/$ – see front matter © 2001 Elsevier Science Ltd. All rights reserved. PII: S0167-7799(01)01671-7

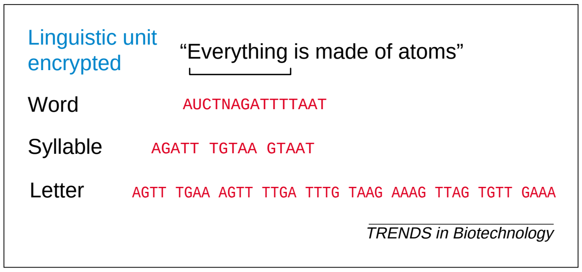
Linguistic unit encrypted
“Everything is made of atoms”
Word
AUCTNAGATTTTAAT
Syllable
AGATT TGTAA GTAAT
Letter
AGTT TGAA AGTT TTGA TTTG TAAG AAAG TTAG TGTT GAAA
TRENDS in Biotechnology

Fig. 1. The first word of Richard Feynman’s suggested message to future civilisations rendered in three different DNA formats.

might be used to encrypt written language because this poses some interesting problems. The principles outlined for encrypting written language, however, also apply to these other types of data.

The first point we should bear in mind when encrypting the message is that we want it to be read. Thus, we should make the encryption system as weak as possible[10]. We should also bear in mind that the receiver would be faced effectively with the need to decipher an ancient script. In the past, successful decipherments (e.g. of the Egyptian hieroglyphs, cuneiform and Linear B) have required large amounts of varied text, parallel translations in one or more known languages and the location of proper names within the script through repetition, highlighting or inspired guesswork[11–14]. When sending our message, we should take these factors into account. A final point to bear in mind is that, in every single case, ancient scripts have been deciphered through a relationship to a known language. However, it ought to be possible to decipher a script of which the language is not known, providing that the language it is written in is placed in context. This can be done by assigning meaning to the words through mathematics, logic and pictures[15], either encrypted in DNA or placed in a key accompanying the message, which should be included in any case.

It probably does not matter much which language we choose for the message – most known languages have a recognizable structure and this structure will be evident on conversion to the base sequence of DNA. There is, of course, no reason why several different languages should not be used and this would probably assist the cryptanalyst.

Having decided on a language, we can proceed to encipher it into the base sequence of DNA. Figure 1 shows how this might be done for letters, syllables and words. In the case of the alphabetic cipher, each letter has been converted to a sequence four nucleotides long. Sticking to genetic nomenclature, we will call this a codon. Note that only three of the four possible bases have been used in the codon and of these G is used only once in any sequence. In their double-stranded forms, therefore, the codons will have isothermal melting temperatures, a point that will become significant in assembling the message. (The idea for using isothermal codons containing only three of the four bases was suggested by Brenner and colleagues[16,17].) The base restrictions within the codons allow 32 permutations, in turn allowing the encipherment of 32 different symbols, enough to cover the letters of the alphabet with some codons left over for other symbols, such as a full stop. One of the symbols could also be used to indicate proper names, in a similar manner to the determinatives of Egyptian hieroglyphics[18].

Using fixed-length codons that employ just three of the four possible bases, and using them in only certain combinations, has several advantages. First, through the patterns generated by the codons, the recipient should be given the strong impression that they are dealing with a message, or at least something very unusual. To expect the recipient to look for linguistic structure in the message DNA, even without prompts left with the collection of amber capsules, is not implausible, and indeed it has been done recently for natural DNA (Ref. 19). Second, the presence of only three types of base in one strand should facilitate the identification of any mutations that occur during the revival of the microorganisms and the concurrent amplification of the message (or during message assembly). Actually, it is possible to devise codon sets in which any single point mutation will be recognised as an error. For example, although they will not be considered further in this article, palindromic codons possess this error-detecting property. Additionally, the four-base codon should distinguish itself from the three base codon of the genetic code. We could have varied the length of the codon according to the frequency that its corresponding symbol appeared in the English language, as in the Morse code, to compress the message and conserve DNA (Refs 20, 21); however, this would complicate the decipherment. A cipher system of 32 codons ought to suggest to the recipient that they are dealing with an alphabetic language, given that such languages have 20–40 symbols (Ref. 22).

Another possibility we might have considered in designing the alphabetic cipher is to have selected codons belonging to a comma-free code[23]. A comma-free code is one in which only one reading frame makes sense and all the others are nonsense. For example, the set of codons AGAA, AGAT, TGAT and TGAA belong to a comma-free code[24] because putting them together in any combination leads to one reading frame, and one reading frame only. A comma-free code would be advantageous if the message became damaged at any point, and the decipherer needed to reorientate themselves in the undamaged sections of the message. However, attempting to combine comma-freedom with the isothermal and base-biased natures of the codons would lead to them becoming very long and would therefore be uneconomical, as well as making it more difficult to spot patterns in the DNA sequence. Given that there will ultimately be multiple copies of the message, through culturing the spores, comma-freedom is probably unnecessary.

Although we have previously spurned an opportunity to compress our message with variable length codons, some compression can be achieved by encrypting words rather than letters (Fig. 1). The codons would then be part of a code rather than a cipher. A vocabulary of 5000 words should be more than adequate for our descriptive needs (Basic English has a vocabulary of only 850 words[25]). To encode this number of words with the same scheme that we used for letters, we would need to increase the length of our codon to ten nucleotides (to give 5120 different permutations). Therefore, for words of three letters or more, the word-encoding system would waste less DNA than the alphabetic cipher. However, if the key that explained the code became damaged in any way, the message would be difficult to decode. Moreover, by encoding words, any inflections in the language, which were so important in the decipherment of Linear B (Refs 11, 14), would be lost. At least with an alphabetic cipher there would be some hope of decryption; alternatively inflection and economy could be combined in a syllabic system (Fig. 1).

Putting the message together

A major challenge presented by this project is the construction of the message. *S. cerevisiae* can stably accommodate 1–2 Mbp of foreign DNA (on a yeast artificial chromosome[26]) and *B. subtilis* can accommodate ~0.1 Mbp (through chromosomal integration[27]). Aiming for chromosome-sized pieces of continuous text, perhaps 10^6 base pairs, a book's worth of material in the alphabetic cipher would be very ambitious but not totally unreasonable in the case of *S. cerevisiae*. (If each spore contained only one 'message chromosome' and 10^8 spores could be stored in one amber capsule, equivalent to 10^8 books, even the storage capacity of one capsule would vastly exceed that of a standard university library, which contains ~10^5 books). At present the largest piece of DNA of defined sequence that can be assembled in the laboratory is ~2000 bp long[28,29]. This was achieved using a DNA polymerase-catalysed reaction in which short overlapping oligonucleotides were allowed to anneal to each other and were then extended by the polymerase to gradually build up the desired segment. A key requirement of this process is that the melting temperatures of the double-stranded overlaps are about the same (i.e. isothermal). By encrypting linguistic units as double-stranded blocks of DNA in which there is one G–C base pair and the others are A–T base pairs we have ensured that this will be the case. The number of errors introduced during the assembly process is relatively small, just less than four point mutations per kbp, certainly not sufficient to jeopardise the sense of the message.

For a 10^6 bp message chromosome 500 of these synthetic 2000 bp fragments would be required. To minimise further errors in the message they could be ligated together via sticky ends, for example, and then further use could be made of the isothermal melting temperatures of the codons. A sticky end of four 'alphabetic' codons, of which there are 32 possibilities for each, could have more than a million (32^4) different manifestations. This greatly exceeds the 499 unique sticky ends that would be required for the isothermal annealing of 500 fragments that are 2000 bp in length, and allows the fragments to be assembled in a desired order, to produce a coherent and not scrambled message. (Annealing at their melting temperature should prevent mis-hybridization of the sticky ends.) Thus, although at first the three-orders-of-magnitude discrepancy between ambition and current reality might have been daunting, provided we monitor the assembly process carefully, building megabase-sized bits of DNA ought to be plausible. The very existence of chromosomes should be encouraging on this point[30,31], although it must be stressed again that the construction work would be a major undertaking.

Orientation of the message

The problem of whether to read the text from left to right or right to left is not an issue with DNA because a message read in one direction could just as easily be read in the other by inverting the double helix. Therefore, one particular end of the message chromosome would have to be designated as the start of the message. This could be done using several repeated sequences to count upwards (repeated sequences at the end of one chromosome and the beginning of another could also be used to align the chromosomes such that the message can be propagated from one chromosome to another).

> '...several major breakthroughs will be required before data retrieval from DNA is as fast as that from conventional storage media.'

Another way that the message orientation could be imparted would be to include successively longer versions of the same message. Taking this paragraph as an example, we could include a fragment encoding the first sentence, another encoding the first two sentences and so on until the entire paragraph was encrypted on one fragment. Then, by aligning repetitions in the fragments, the orientation would become clear. With the alphabetic or syllabic cipher, the orientation of the message should also be apparent from any inflections in the language.

Complications

There are several problems with the proposed scheme. The obvious ones are that: (1) the message is somehow obliterated, deliberately or otherwise; (2) we become extinct and therefore there is no-one to receive the message; (3) we evolve into species that have different linguistic capabilities to ourselves; and (4) spore revival is poor, generating an incomplete message. Less obviously, there is also the problem of a hoax message. This would clearly involve a lot of effort but this has not stopped people producing forgeries in the past (as a composite bird fossil[32] and the photograph of an Indonesian coelacanth[33–35] have recently shown). However, in these instances there was presumably a strong motive (money and/or prestige); with a fake message, the only motive would be mischievousness or perhaps religious fervour. Another problem, which ironically would act as a powerful deterrent to attempted forgery, is the cost of producing the message. At present, the oligonucleotides required for one synthetic chromosome would cost more than a million pounds sterling. Finally, the intended recipients might get the message but not realise that it is a message and label it as junk or parasitic DNA (Ref. 36), as we have done when faced with odd genomic features, such as repetitive sequences in the human genome and the dispensable 2μ plasmid of *S. cerevisiae*.

Conclusion

Although DNA is a superb material for storing data, there are two problems that, for the time being, make harnessing its potential difficult. First, retrieving the data from the nucleotide sequence is cumbersome. With improvements in sequencing technologies (e.g. the use of ion channels[37] or atomic force microscopy) this situation is likely to improve, although several major breakthroughs will be required before data retrieval from DNA is as fast as that from conventional storage media. Second, constructing large pieces of synthetic DNA (>10 kbp) is likely to be technically demanding and extremely expensive. However, the idea of assembling sticky ended DNA fragments outlined above could be tested immediately with any large DNA fragments (>1 kbp) that are to hand in the laboratory. Similarly, inventing economical and practical DNA encryption systems that generate obviously artificial base patterns is something anyone can apply their ingenuity to given a pencil, some paper and a bit of thought.

Acknowledgements
I thank Andrew Doig, Pooja Kumar and Dudley Williams who kindly read the manuscript and also Geoff Smith, Colin Harwood, Ian Williams and Alan Wheals for valuable discussions.

References

1 Hoch, J.A. and Losick, R. (1997) Panspermia, spores and the *Bacillus subtilis* genome. *Nature* 390, 237–238
2 Clelland, C.T. *et al.* (1999) Hiding messages in DNA microdots. *Nature* 399, 533–534
3 Anon. (2000) A Y3K bug. *Nat. Biotechnol.* 18, 1
4 Ryder, O.A. *et al.* (2000) DNA banks for endangered animal species. *Science* 288, 275–277
5 Morin, P.A. *et al.* (2000) Preservation of DNA from endangered species. *Science* 289, 725–727
6 Stokstad, E. (2000) Feathers, or flight of fancy? *Science* 288, 2124–2125
7 Nicholson, W.L. *et al.* (2000) Resistance of *Bacillus* endospores to extreme terrestrial and extraterrestrial environments. *Micobiol. Mol. Biol. Rev.* 64, 548–572
8 Cano, R.J. and Borucki, M.K. (1995) Revival and identification of bacterial spores in 25- to 40-million-year-old Dominican amber. *Science* 268, 1060–1064
9 Vreeland, R.H. *et al.* (2000) Isolation of a 250 million-year-old halotolerant bacterium from a primary salt crystal. *Nature* 407, 897–900
10 Kahn, D. (1996) Messages from outer space. In *The Codebreakers*, p. 952, Scribner, New York, USA
11 Singh, S. (1999) The language barrier. In *The Code Book*, pp. 191–242, Fourth Estate, London, UK
12 Bermant, C. and Weitzman, M. (1979) Cuneiform without tears. In *Ebla: An Archaeological Enigma*, pp. 70–123, Weidenfeld and Nicolson, London, UK
13 Pope, M. (1999) Persian cuneiform. In *The Story of Decipherment*, p. 95, Thames and Hudson, London, UK
14 Chadwick, J. (1967) Birth of a theory. In *The Decipherment of Linear B*, pp. 40–66, Cambridge University Press, London, UK
15 Sagan, C. (1997) *Contact*, pp. 233–236, Orbit, London, UK
16 Brenner, S. *et al.* (2000) *In vitro* cloning of complex mixtures of DNA on microbeads: physical separation of differentially expressed cDNAs. *Proc. Natl. Acad. Sci. U. S. A.* 97, 1665–1670
17 Brenner, S. *et al.* (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat. Biotechnol.* 18, 630–634
18 Davies, W.V. (2000) The principles. In *Egyptian Hieroglyphs*, pp. 33–36, British Museum Press, London, UK
19 Tsonis, A.A. *et al.* (1997) Is DNA a language? *J. Theor. Biol.* 184, 25–29
20 Kahn, D. (1996) The anatomy of cryptology. In *The Codebreakers*, p. 741, Scribner, New York, USA
21 Doig, A.J. (1997) Improving the efficiency of the genetic code by varying the codon length – the perfect genetic code. *J. Theor. Biol.* 188, 355–360
22 Singh, S. (1999) The language barrier. In *The Code Book*, p. 221, Fourth Estate, London, UK
23 Crick, F.H.C. *et al.* (1957) Codes without commas. *Proc. Natl. Acad. Sci. U. S. A.* 43, 416–421
24 Golomb, S.W. (1962) Efficient coding for the desoxyribonucleic channel. *Proceedings of the Symposium for Applied Mathematics* 14, 87–100
25 Ogden, C.K. (1968) *Basic English: International Second Language*, p. 5, Harcourt, Brace & World, New York, USA
26 Burke, D.T. *et al.* (1987) Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors. *Science* 236, 806–812
27 Harwood, C.R. (1992) *Bacillus subtilis* and its relatives: molecular biological and industrial workhorses. *Trends Biotechnol.* 10, 247–256
28 Withers-Martinez, C. *et al.* (1999) PCR-based gene synthesis as an efficient approach for expression of the A+T-rich malaria genome. *Protein Eng.* 12, 1113–1120
29 Stemmer, W.P.C. *et al.* (1995) Single-step assembly of a gene and entire plasmid from large numbers of oligodeoxyribonucleotides. *Gene* 164, 49–53
30 Maynard Smith, J. and Szathmáry, E (1993) The origin of chromosomes I. Selection for linkage. *J. Theor. Biol.* 164, 437–446
31 Szathmáry, E. and Maynard Smith, J. (1993) The evolution of chromosomes II. Molecular mechanisms. *J. Theor. Biol.* 164, 447–454
32 Rowe, T. *et al.* (2001) The *Archaeoraptor* forgery. *Nature* 410, 539–540
33 McCabe, H. and Wright, J. (2000) Tangled tale of a lost, stolen and disputed coelacanth. *Nature* 406, 114
34 McCabe, H. (2000) Recriminations and confusion over 'fake' coelacanth photo. *Nature* 406, 225
35 Erdmann, M.V. and Caldwell, R.L. (2000) How new technology put a coelacanth among the heirs of Piltdown Man. *Nature* 406, 343
36 Orgel, L.E. and Crick, F.H.C. (1980) Selfish DNA: the ultimate parasite. *Nature* 284, 604–607
37 Vercoutere, W. *et al.* (2001) Rapid discrimination among individual DNA hairpin molecules at single-nucleotide resolution using an ion channel. *Nat. Biotechnol.* 19, 248–252

Do you wish to contribute an article to *TIBTECH*?

If so, send a brief (half to one page) outline of the proposed content of your article, stating which section of the journal you wish it be considered for.

You may also suggest topics and issues that you would like to see covered by the journal.

Please contact:
The Editor, *Trends in Biotechnology*,
Elsevier Science London,
84 Theobald's Road, London,
UK WC1X 8RR.
(Fax: +44 (0)20 7611 4470)