Artificial DNA: PNA & XNA 3:3, 109–111; July–December 2012; © 2012 Landes Bioscience

To DNA, all information is equal

Lau Sennels[1,*] and Thomas Bentin[2]

[1]Institute of Cell Biology; University of Edinburgh; Edinburgh, UK; [2]Department of Cellular and Molecular Medicine; The Panum Institute; University of Copenhagen; Copenhagen, Denmark

Keywords: DNA, information storage in DNA, bit, byte, binary encoding

Submitted: 10/23/12

Accepted: 10/25/12

http://dx.doi.org/10.4161/adna.22671

*Correspondence to: Lau Sennels;
Email: lau@sennels.dk

Comment on: Church GM, Gao Y, Kosuri S. Next-generation digital information storage in DNA. Science 2012; 337:1628; PMID:22903519; http://dx.doi.org/10.1126/science.1226355

Information storage capabilities are key in most aspects of society and the requirement for storage capacity is rapidly expanding. In principle, DNA could be a high-density medium for information storage. Church and coworkers recently demonstrated how binary data can be encoded, stored in and retrieved from a library of oligonucleotides, increasing by several orders of magnitude the amount and density of manmade information stored in DNA to date. The technology remains in its infancy and important hurdles have yet to be overcome in order to realize its potential. However, DNA may be particularly useful as a storage-medium over long time-scales (centuries), because data-access is compatible with any large-scale DNA-sequencing and -synthesis technology.

It has been proposed that DNA could be a storage-medium for non-biological (man-made) information.[1] In DNA, information is encoded directly in the primary structure as chemically well-defined monomers, the nucleobases, linked together in linear arbitrary-length heteropolymers. In physiological conditions, DNA consists of two self-complementary strands maintaining information-integrity by facilitating template-based, enzyme-driven error-correction. Furthermore DNA molecules have considerable physicochemical stability with presumed lifetimes on the order of centuries.[1-3] Given a suitable scheme for encoding arbitrary information as nucleobase sequences, DNA-storage could consequently be an attractive storage-medium, with theoretical storage-densities in the exabyte (10^{18} bytes)/mm^3-range, well beyond what is currently achievable using magnetic or optical storage.

In practice, however, neither the theoretical storage-density nor stored information amounts have been achieved. A number of works have successfully encoded human-readable text as complimentary DNA-oligonucleotides, inserted into self-replicating vectors maintained in bacteria, or inserted into the host genome itself. The potential of this approach is perhaps most radically exemplified by the synthesis of the entire 1.1 MBp chromosome of *Mycoplasma mycoides* assembled (using PCR and baker's yeast) from chemically synthesized oligonucleotides and subsequently introduced via "transplantation" into a related species.[4] In principle, such engineering allows information-encoding on a MBit-scale, even though the technology was not developed explicitly for storage-purposes. However, the technical difficulty and cost of synthesizing long, error-free DNA-sequences on a large scale, and the need for host cell transformation, sets limits for the stored information amount and density. Additionally, irrespective of how information is stored in DNA, a significant obstacle to any current, practical use of DNA-storage technology, is the time and cost associated with DNA-sequencing for high-capacity information retrieval.

Church and colleagues[3] now demonstrate how binary data can be encoded and written as an in vitro library of synthetic DNA oligonucleotides, and subsequently read using next-generation sequencing to reconstruct the information. Using this technique the authors write an e-book of 650 kB (about 2.5 times the size of the PDF of their paper) to DNA-storage

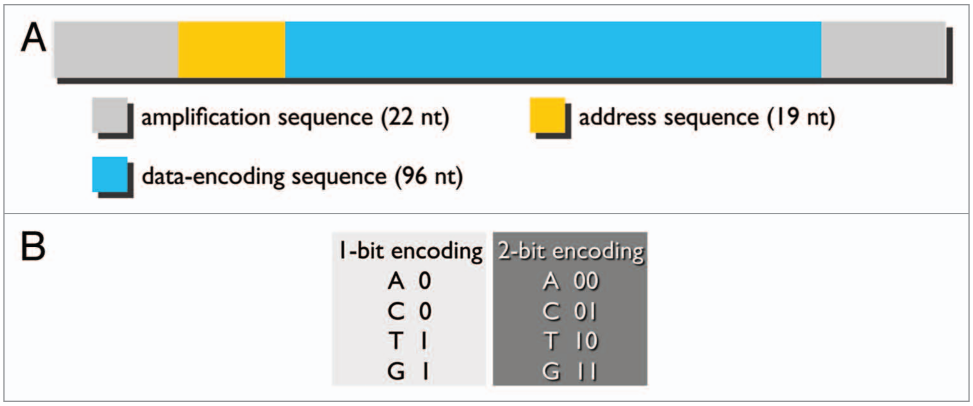

Figure 1. (**A**) Structure of DNA information storage space and (**B**) binary encoding scheme.

medium, and demonstrate its retrieval. Even if the amount of information stored is negligible compared with the routine capacity of conventional storage, it still is a 670-times improvement over previous work in biological systems. Furthermore, the data-density is approximately 100 times higher than previous works employing DNA, and 6 orders-of-magnitude higher than conventional hard drives, according to the authors.[3]

Rather than storing the approximately 658,000-"character"-sequence in one contiguous DNA-sequence, Church and colleagues[3] constructed a library of 54.898 159 nt ssDNA-oligonucleotides, each containing the DNA-encoded form of up to 12-character contiguous fragments from the original sequence (96 nt, and see below), a unique address-sequence (19 nt), and common 5' and 3' primer binding sites for library amplification and sequencing (**Fig. 1A**). By splitting the encoding DNA-sequences into short amplifiable vectors, the library can be chemically synthesized in a high-throughput manner amplified by PCR and subsequently sequenced for retrieval of the encoded information source.

In order to demonstrate the ability to store the same type of information as conventional digital storage, the authors[3] chose to encode an eBook, containing English text, HTML (the WWW markup language for documents), images, as well as a short piece of program code, executable by most web browsers. Together, these items constitute most of the different types of information-representations commonly manipulated by computers.

By representing text and images as a binary sequence, the eBook was encoded by a minimal DNA-sequence.[3] Text strings of the eBook were encoded in UTF-8 (universal character set transformationformat-8 bit). UTF-8 is a format representing characters as multiples of 8 bit (1 byte) in binary numbers, providing universally standardized encoding for any character with a computer-representation in Unicode.[5] With four canonical nucleobases, DNA can maximally encode two bits/base (**Fig. 1B**). However, Church and colleagues[3] opted for a 1-bit representation, allowing a given position in the binary sequence to be represented in two ways in the DNA-sequence. As stated by the authors, this redundant coding scheme made it possible to balance GC content and avoid homopolymeric runs, e.g., posing difficulty during readout. Furthermore, it adds robustness against information-corruption, since transversions from A to C, or T to G are "silent" with respect to the integrity of the encoded information.[3] To this end, we note that the coding scheme remains sensitive toward A to T and T to A transversions, as well as the, in Nature, more frequently occurring transitions.

In the representation proposed by Church and colleagues,[3] the source information is stored in blocks of 96 bp, (96 bits or 12 bytes), preceded by a unique 16 bp address tag, so that the original source can be reassembled from the sequenced blocks (**Fig. 1A**). Consequently, the theoretical maximum storage capacity is 2^{19} times 12 bytes or 6.3 MB, for the particular choice of length of address tag and storage sequence.

Church and colleagues[3] demonstrate an in vitro DNA-encoding storage scheme for digital information, and suggest its suitability, in principle, for large-scale storage. In this scheme, the theoretical maximum storage-capacity is constrained by the size of the address and data-encoding sequences. Extending those sequences would increase storage-capacity. For example, increasing the length of the address-tag to 48 bp (48 bit), would increase the theoretical capacity into the petabyte (10^{15}) range. However, the current practical storage-capacity is limited by the ssDNA-library synthesis capacity. That is, constrained by the ssDNA-oligonucleotide sequence length and throughput (number) of oligonucleotides, which can reliable be synthesized, and ultimately by the maximum length that can be consistently covered by paired-end sequencing. The eBook was stored as a ssDNA-library of 54,898 oligonucleotides, about 6% of the largest feasible synthesis capacity for current microarrays,[6] and about 10% of the capacity available with a 19 bit address-tag. Consequently, it would be feasible to encode an information amount corresponding to the full capacity of the address-space (6.1 MB), but not practical to encode amounts on par with the storage capacity of standard hard drives, which typically come in the range of 500 GB–2 TB at present. The stored information was retrieved with high coverage of the ssDNA-library. The average sequencing depth for a given ssDNA-vector was 3,419 ± 998 times, with approximately 70% of reads being full-length. However, when selecting only vectors where the address tag was canonical, the tail 2% (estimated) had less than 100 copies, some less than 10.[3] This illustrates the limitation imposed by sequencing-depth on data-retrieval. Merely doubling the ssDNA-library, corresponding to a stored amount of 12.2 MB, would reduce the copy number of low-coverage vectors disproportionally. As a consequence some vectors might not be represented at all, and sequencing-errors in low-coverage vectors would have proportionally greater probability of resulting in information-corruption. Furthermore, 11% of full-length reads had non-canonical address-tags.[3] This could result in corruption of the reassembled information source where entire data-blocks (12 bytes) might be shuffled or deleted.

DNA-storage, as implemented by Church and colleagues, could conceivably be practical for large-scale applications where storage-times would be very long (on the order of centuries) and information-access rare.[3] The high storage-density and relative stability of DNA could make it an alternative to current mechanical or optical storage devices for such applications, which would have significant space and power requirements for retaining information-integrity over long time-scales. The stipulation of rare access-times is important because data access-times are much longer for DNA-storage, compared with hard drives: days vs. milliseconds, respectively.

For the prospect of large-scale DNA-storage be realistic, significant gains in scale of oligonucleotide-synthesis and DNA-sequencing would have to be made. The authors estimate that an increase in scale of synthesis on the order of 7-orders-of-magnitude, and a corresponding 6-orders-of-magnitude increase of scale of sequenced features would be necessary to reach storage-capacities in the exabyte-range.[3]

We think the most intriguing argument for the use of century-long DNA-storage over other technologies is that it is comparatively implementation-neutral. Since information is stored in the primary structure of DNA, any technology capable of synthesizing or sequencing DNAs could conceivably be used to write or read stored information, respectively. The same is not the case for, e.g., a hard drive. The information stored on a hard drive's magnetic disk can only be accessed safely through the electronics of the device, and requires specific knowledge of its hardware and software interface. In a century, computer-technology will have developed very significantly, and will undoubtedly be incompatible with hardware and software developed in the present. If information stored today were to be accessed in a century, the required expertise, software and hardware for doing so will likely have been lost. For a contemporary example of this problem, consider the difficulty associated with accessing data on a HD 3.5"-floppydisk from a standard PC, even though the medium became available only 25 years ago, and remained in widespread use into this century. For information stored today or in the near future for the benefit of our descendents in centuries or millennia to come, ensuring media-compatibility, will be paramount.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

References

1. Bancroft C, Bowler T, Bloom B, Clelland CT. Long-term storage of information in DNA. Science 2001; 293:1763-5; PMID:11556362; http://dx.doi.org/10.1126/science.293.5536.1763c.
2. Bonnet J, Colotte M, Coudy D, Couallier V, Portier J, Morin B, et al. Chain and conformation stability of solid-state DNA: implications for room temperature storage. Nucleic Acids Res 2010; 38:1531-46; PMID:19969539; http://dx.doi.org/10.1093/nar/gkp1060.
3. Church GM, Gao Y, Kosuri S. Next-generation digital information storage in DNA. Science 2012; 337:1628; PMID:22903519; http://dx.doi.org/10.1126/science.1226355.
4. Gibson DG, Glass JI, Lartigue C, Noskov VN, Chuang RY, Algire MA, et al. Creation of a bacterial cell controlled by a chemically synthesized genome. Science 2010; 329:52-6; PMID:20488990; http://dx.doi.org/10.1126/science.1190719.
5. The Unicode Consortium. The Unicode Standard, Version 6.2.0. (2012), http://www.unicode.org/versions/Unicode6.2.0/
6. LeProust EM, Peck BJ, Spirin K, McCuen HB, Moore B, Namsaraev E, et al. Synthesis of high-quality libraries of long (150mer) oligonucleotides by a novel depurination controlled process. Nucleic Acids Res 2010; 38:2522-40; PMID:20308161; http://dx.doi.org/10.1093/nar/gkq163.