Long-Term Memory

DOI: 10.1002/anie.201411378

Robust Chemical Preservation of Digital Information on DNA in Silica with Error-Correcting Codes**

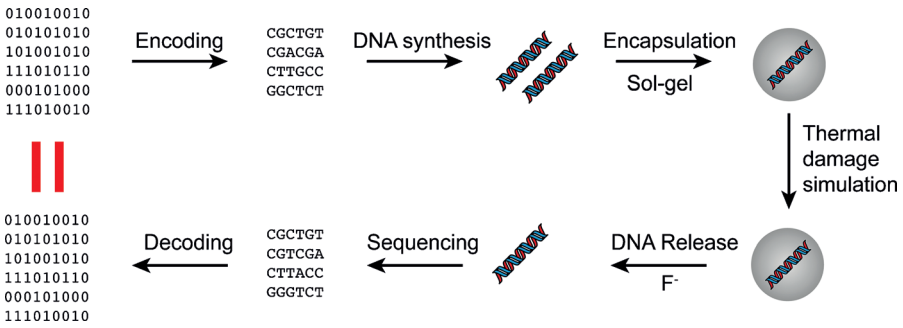
Robert N. Grass, Reinhard Heckel, Michela Puddu, Daniela Paunescu, and Wendelin J. Stark*

Abstract: *Information, such as text printed on paper or images projected onto microfilm, can survive for over 500 years. However, the storage of digital information for time frames exceeding 50 years is challenging. Here we show that digital information can be stored on DNA and recovered without errors for considerably longer time frames. To allow for the perfect recovery of the information, we encapsulate the DNA in an inorganic matrix, and employ error-correcting codes to correct storage-related errors. Specifically, we translated 83 kB of information to 4991 DNA segments, each 158 nucleotides long, which were encapsulated in silica. Accelerated aging experiments were performed to measure DNA decay kinetics, which show that data can be archived on DNA for millennia under a wide range of conditions. The original information could be recovered error free, even after treating the DNA in silica at 70 °C for one week. This is thermally equivalent to storing information on DNA in central Europe for 2000 years.*

Prehistorical information put down by our ancestors in cave drawings, texts engraved in gold, and medieval texts are some of the strongest links with our past. An example is the Archimedes Palimpsest that originates from the tenth century. This contains the single known copy of "The Methods of Mechanical Theorems", and represents a cornerstone in the development of geometry and modern calculus. The book has survived more than 1000 years and in 1998 was valued at more than two million USD. In view of this valuation of information it may seem surprising that current efforts of guaranteeing longevity of digital information are scarce (e.g. M-Disc, Syylex) and the storage half-life of information has dropped drastically since the transition from analog to digital storage systems.[1]

Traditional storage technologies such as optical and magnetic devices are not reliable for long-term (> 50 years) data storage.[2] Furthermore, the development of reliable systems requires long-term testing, which is well above the current device-development timelines. DNA is the only data-storage medium for which real long-term data are available from archeology. Most recently, 300 000 year old mitochondrial DNA from bears and humans has been sequenced.[3] DNA has also previously been utilized as a coding language, for applications in forensics,[4] product tagging,[5] and DNA computing.[6] As a consequence, several approaches to store information on DNA have been proposed.[7] However, those approaches are not reliable as they cannot handle errors efficiently and do not suggest how to (physically) store the DNA to maintain its stability over time.

To overcome these issues we combined an error-correcting information-encoding scheme tailored to DNA (Scheme 1) with a previously established chemical method for storing DNA in "synthetic fossils". The corresponding experiments show that only by the combination of the two concepts, could digital information be recovered from DNA stored at the Global Seed Vault (at −18 °C) after over 1 million years.



Scheme 1. Digital information is encoded to DNA and encapsulated within silica spheres. Upon release of the DNA from the spheres by fluoride chemistry, the DNA is read by Illumina sequencing and decoded to recover the original information, even if errors were introduced during the procedures.

[*] Dr. R. N. Grass, M. Sc. M. Puddu, M. Sc. D. Paunescu, Prof. W. J. Stark
Institute for Chemical and Bioengineering
ETH Zurich
Vladimir-Prelog-Weg 1, 8093 Zurich (Switzerland)
E-mail: robert.grass@chem.ethz.ch
Homepage: www.fml.ethz.ch
Dr. R. Heckel
Department of Information Technology and Electrical Engineering
ETH Zurich
Sternwartstrasse 8, 8092 Zurich (Switzerland)

[**] We would like to thank the Institute of Chemical and Bioengineering of ETH Zurich, the Swiss National Science Foundation grant (no. 200021-150179), and the EU-ITN network Mag(net)icFun (PITN-GA-2012-290248) for financial support. We thank Christof Wunderlin (Microsynth AG) and Marcello Caraballo (Customarray Inc.) for support with DNA synthesis and sequencing.

Supporting information for this article is available on the WWW under http://dx.doi.org/10.1002/anie.201411378.

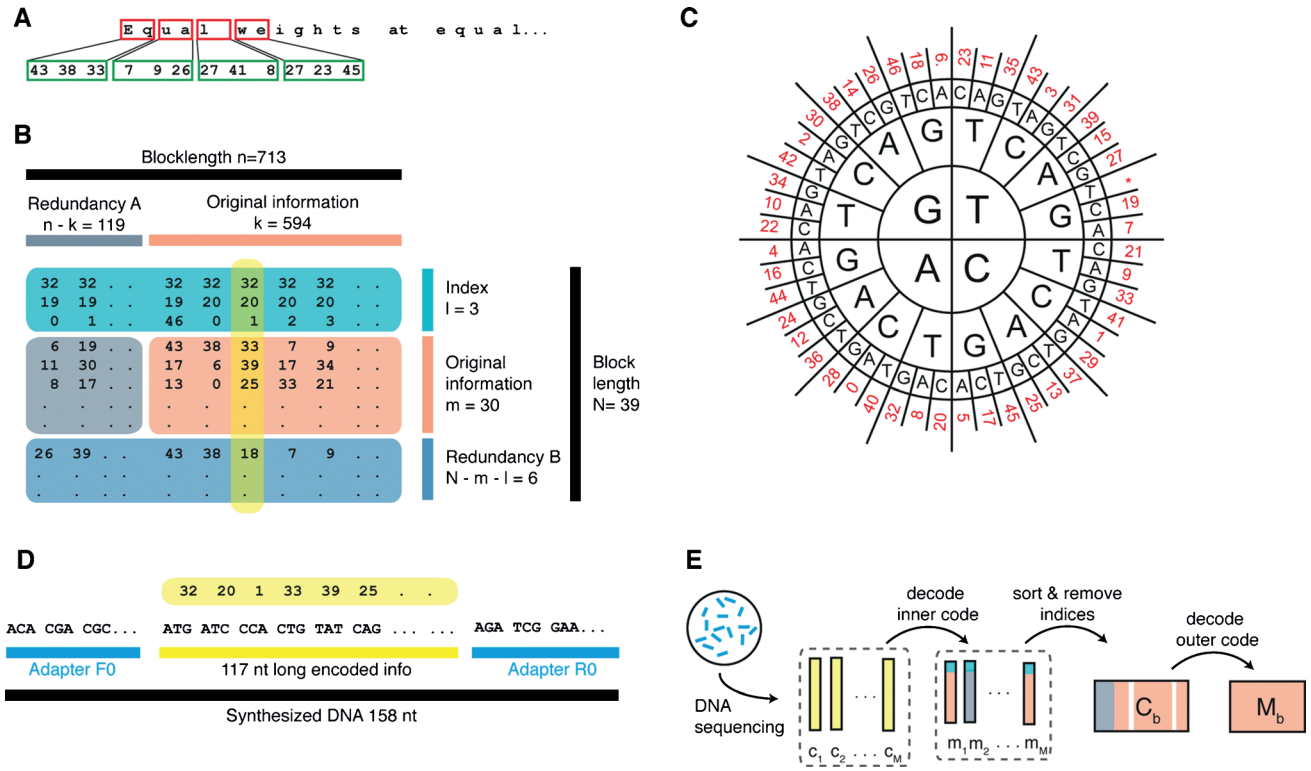


Figure 1. Encoding text to DNA by Reed–Solomon coding: A) Two letters of a text file (or more general, two bytes of a digital file) are mapped to three elements of the Galois Field of size 47 (GF(47)) by base conversion (256^2 to 47^3). This original information is arranged in blocks of 594×39 elements. B) In an outer encoding step Reed–Solomon (RS) codes are employed to add redundancy A to the individual blocks. To each column an index is added and redundancy B is generated using a second (inner) RS encoding step. C) The individual columns are converted into DNA by mapping every element of GF(47) to three nucleotides by utilizing the GF(47)toDNA codon wheel, thereby guaranteeing that no base is repeated more than three times. D) Two constant adapters are added and the resulting sequences of 158 nucleotides are synthesized. E) To recover the original information from the DNA, the read sequences are translated to GF(47) and are decoded by first decoding the inner code (correcting individual base errors), sorting the sequences by means of the index, followed by outer-decoding, which allows the correction of whole sequences and the recovery of completely lost sequences (see the Supporting Information for details on coding and experimental procedures).

Since synthesis and sequencing of very long DNA strands is technically impeded, data must be stored on several short DNA segments, which cannot be arranged geometrically. Writing and reading DNA is, therefore, different to commonly used storage technologies such as magnetic disks. Moreover, DNA-specific errors are expected during writing, long-term storage, and DNA reading (sequencing). These are the loss of individual base information as well as the loss of complete sequences. In classical data-storage devices, error-correcting codes are implemented, which add redundancy and allow the correction of essentially all errors that occur during usage. To account for the specific requirements of storage on DNA the existing data coding schemes had to be adapted: Individual sequences are indexed and two independent error-correcting codes (specifically Reed–Solomon codes)[8] are used in a concatenated fashion (Figure 1; see the Supporting Information for the rationale of code design and parameter choice).

To physically test the code we stored the text from two old documents: the Swiss Federal Charter from 1291 and the English translation of the Method of Archimedes. The (uncompressed) total text is 83 kilobytes large, and was encoded as shown in Figure 1. This resulted in 4991 sequences, each 117 nucleotides long to which constant primers were added (giving a total length of 158 nt) to allow for a rapid and indexed library preparation for sequencing. The sequences were synthesized on an electrochemical microarray technology (CustomArray),[9] prepared for sequencing by a custom PCR (polymerase chain reaction) method, and read using the Illumina MiSeq platform (see the Supporting Information for experimental details). From reading the sequences, the inner code had to correct an average of 0.7 nt errors per sequence and the outer code had to account for a loss of 0.3% of total sequences and correct about 0.4% of the sequences, thereby resulting in a complete and error-free recovery of the original information.

This experiment demonstrates that information can be stored on DNA reliably. It still remains to be shown if DNA can indeed be utilized for ultralong storage times, as DNA in solution decays within several years.[10] To test if DNA stored in the solid state is more stable,[11] we took the 4991 element oligo pool and tested the stability of three previously established dry storage procedures for DNA by accelerated aging tests (Figure 2). The individual technologies are a storage format on impregnated filter paper,[12] a biopolymer technology that mimics the anhydrous vitreous state of DNA

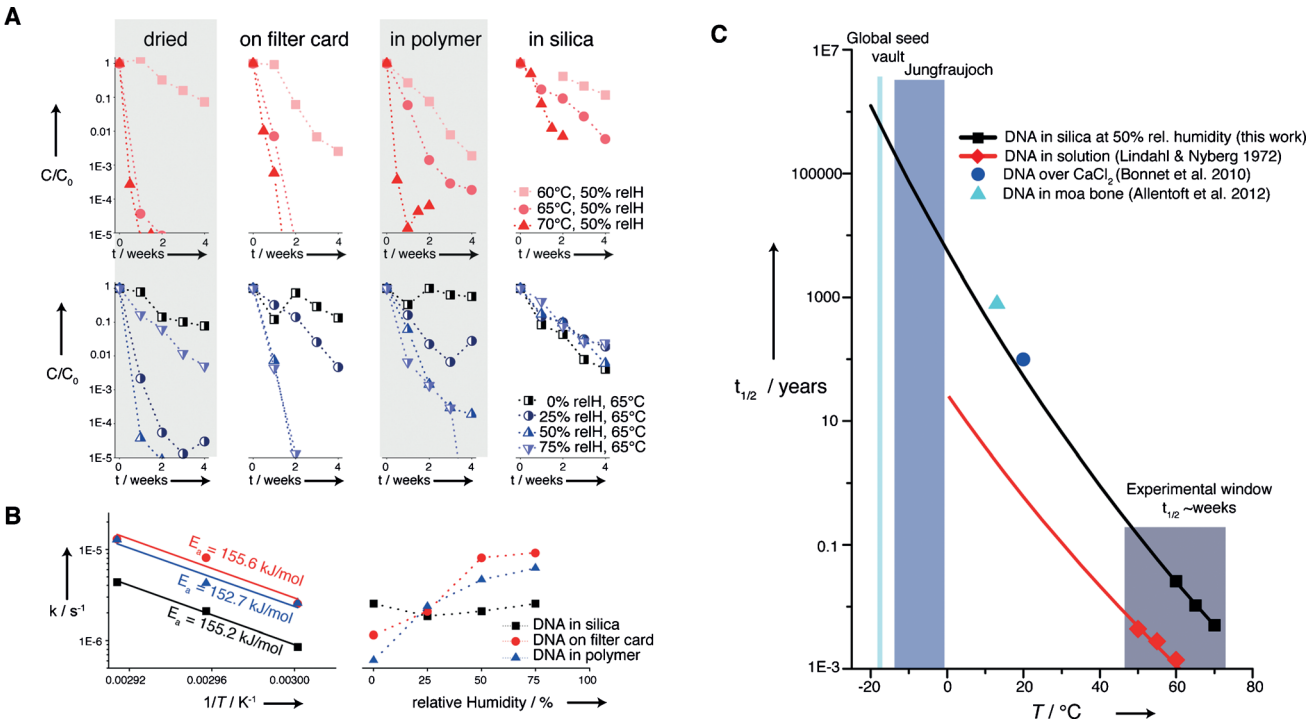


Figure 2. Degradation kinetics of dry DNA storage: A) Effect of temperature and humidity on DNA integrity (per qPCR) as a function of time using four different dry DNA storage technologies: pure solid-state DNA, DNA on FTA filter cards, DNA in a biopolymeric DNA storage matrix (DNAstable), and DNA encapsulated in silica (data at 70 °C were collected at a higher rate to account for the accelerated kinetics). B) First-order decay rate constants derived thereof and activation energy of degradation processes assuming first-order kinetics. C) Half-life of 158 bp DNA stored in a silica matrix extrapolated according to the Arrhenius Equation with activation energies of 155 ± 10 kJ mol^{-1} and compared to literature data on DNA stability in solution,[10] desiccated DNA stability,[11] and DNA stability in ancient moa bone.[15] Data from the literature are scaled[15] by $t_{1/2}{}^{158\,nt} = t_{1/2}{}^{1\,nt}/158$.

in seeds and spores,[13] and a synthetic silica fossilization technology based on a procedure developed in our group.[14] Compared to the storage of solid-state DNA without additional agents, all three solid-state DNA storage technologies decreased the DNA decay rates considerably. From the temperature dependence of the decay rates, Arrhenius-type activation energies (E_A) were calculated by assuming first order kinetics, which were equivalent for all three storage formats (155 ± 2 kJ mol^{-1}; see the Supporting Information for details). This is in line with recent data on the kinetics of single-strand breaks in dry DNA storage ($E_A = 158$ kJ mol^{-1})[11] and differs considerably from the previously established activation energy of DNA depurination in solution (105–120 kJ mol^{-1}).[10] Although the activation energy is near to identical for the three storage formats, the individual decay rates differ. As a result of the humidity dependence, the decay kinetics of solid-state DNA can be best represented by Equation (1):

$$\frac{dc_{DNA}}{dt} = k_0 \cdot \left(c_{H_2O}\right)^n \cdot e^{\frac{-E_A}{RT}} \cdot c_{DNA} = A \cdot e^{\frac{-E_A}{RT}} \cdot c_{DNA} \qquad (1)$$

where the observed factor A accounts for the preexponential Arrhenius factor k_0 and the effect of water $\left(c_{H_2O}\right)^n$. It may be concluded from the identical activation energies that DNA degrades by the same single-strand break mechanism,[11,16] and the individual decay rates depend merely on the storage temperature and the water concentration within the vicinity of the DNA molecules. (It is to be expected that water is associated with DNA when stored within biopolymers and even when encapsulated within silica.) From the data shown in Figure 2 it is evident that DNA preservation is best in the inorganic storage format (DNA encapsulated in silica), which has the lowest local water concentration. By separating the DNA molecules from the environment by an inorganic layer, the degree of preservation is not affected by the humidity of the storage environment. This independence of humidity is very important for guaranteeing long-term stability, as a non-humid environment is hard to maintain. In contrast, stability-increasing factors such as low temperature (e.g. permafrost) and absence of light can be maintained for extended periods of time without energy input. The DNA storage system within silica further offers exceptional stability against oxidation (see the reactive oxygen species (ROS) test in the Supporting Information), and by adding an additional titania layer the photoresistance of DNA can be greatly improved.[17]

In ancient fossil bone, DNA has the greatest chance of survival if encapsulated within apatite/collagen structures[18] and crystal aggregates,[19] which protect the solid DNA from the environment and humidity—very similar to the encapsulation of DNA within the inorganic silica particles utilized here. Indeed if the decay kinetics presented in Figure 2 for DNA encapsulated in silica are extrapolated to lower temperatures, the data coincides well with the degradation

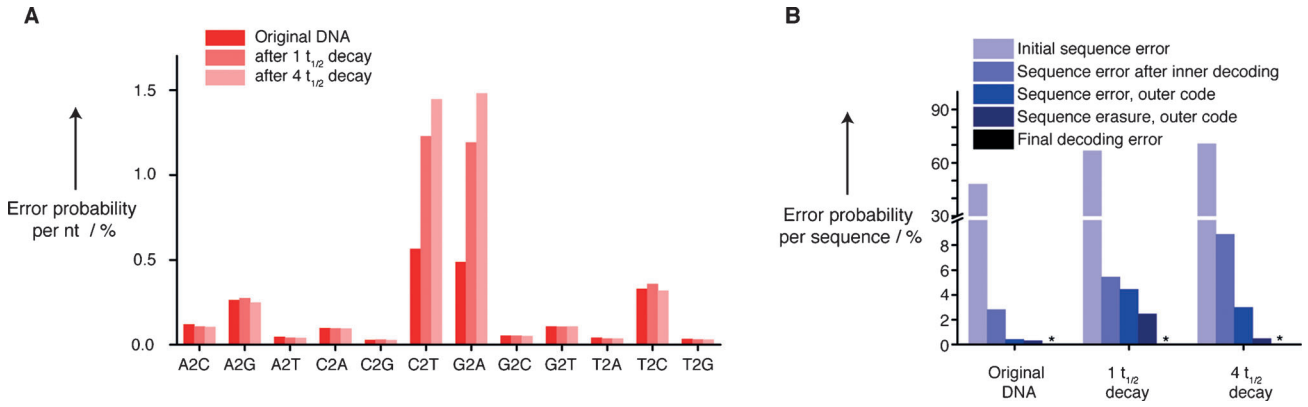


Figure 3. Decoding statistics and error correction: A) Error probabilities of individual base calls (e.g. A2C is the probability of reading C instead of A). B) Probability of a sequence to contain errors (initial error), as well as sequence errors and erasures handled by the inner and outer decoder resulting in a final decoding error probability of 0 (*) for all cases (see the Supporting Information for definitions and discussion of errors).

rate of DNA in fossil moa bone derived by Allentoft et al. from samples up to 8000 years old.[15] Additionally, the decay rates coincide with previously reported decay rates on desiccated DNA storage over the course of 32 weeks (point 4 in Bonnet et al.[11]). Sufficient stability to explain sequencing success from 300000 year old bone samples is also given (see discussion in the Supporting Information). This indicates that in both cases (DNA in bone and DNA in silica) the decay of information follows the same kinetic rate law and the accelerated aging experiments performed with DNA in silica mimic the long-term decay of DNA in fossil bone.

To show that the information stored in the synthetic DNA can still be read after significant thermal treatment, DNA stored within silica particles at 70 °C for half and for one week was sequenced. The decoding scheme presented in Figure 1 was utilized to reconstruct the data. These two data points represent DNA degraded by about one and four decay half-lifes. Although the inner and outer code of the error-correcting scheme had to correct significantly more errors than in the non-heat-treated sample, in both cases the original information could be recovered without final error (Figure 3).

Being able to reconstruct the original data from DNA decayed for 4 half-lifes, which according to Figure 2c is equivalent to storing DNA in Zurich (9.4 °C) for 2000 years, or at the coldest, year-round accessible place in Switzerland (Jungfraujoch, 3471 m above sea level) for about 100000 years. These data further predict that digital information could be stored encapsulated in silica at the Global Seed Vault (at −18 °C) for over 2 million years.

Received: November 24, 2014
Published online: February 4, 2015

Keywords: DNA · fossils · information storage · long-term memory · sol-gel processes

[1] a) M. Hilbert, P. Lopez, *Science* **2011**, *332*, 60; b) P. Conway, *Libr. Q.* **2010**, *80*, 61.
[2] S. Shah, J. G. Elerath, *Annu. Reliab. Maintainability Symp. Proc.* **2004**, 163.
[3] a) J. Dabney et al., *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 15758; b) M. Meyer et al., *Nature* **2014**, *505*, 403.
[4] J. M. Oh, D. H. Park, J. H. Choy, *Chem. Soc. Rev.* **2011**, *40*, 583.
[5] D. H. Park, C. J. Han, Y. G. Shul, J. H. Choy, *Sci. Rep.* **2014**, *4*, 4879.
[6] a) Z. Ezziane, *Nanotechnology* **2006**, *17*, R27; b) Y. Benenson, *Nat. Rev. Genet.* **2012**, *13*, 455.
[7] a) C. Bancroft, T. Bowler, B. Bloom, C. T. Clelland, *Science* **2001**, *293*, 1763; b) G. M. Church, Y. Gao, S. Kosuri, *Science* **2012**, *337*, 1628; c) N. Goldman, P. Bertone, S. Y. Chen, C. Dessimoz, E. M. LeProust, B. Sipos, E. Birney, *Nature* **2013**, *494*, 77; d) J. Davis, *Art J.* **1996**, *55*, 70.
[8] I. S. Reed, G. Solomon, *J. Soc. Ind. Appl. Math.* **1960**, *8*, 300.
[9] S. Kosuri, G. M. Church, *Nat. Methods* **2014**, *11*, 499.
[10] T. Lindahl, B. Nyberg, *Biochemistry* **1972**, *11*, 3610.
[11] J. Bonnet, M. Colotte, D. Coudy, V. Couallier, J. Portier, B. Morin, S. Tuffet, *Nucleic Acids Res.* **2010**, *38*, 1531.
[12] L. A. Burgoyne, US Patent 6322983B, **2001**.
[13] E. Wan, M. Akana, J. Pons, J. Chen, S. Musone, P. Y. Kwok, W. Liao, *Curr. Issues Mol. Biol.* **2010**, *12*, 135.
[14] a) D. Paunescu, R. Fuhrer, R. N. Grass, *Angew. Chem. Int. Ed.* **2013**, *52*, 4269; *Angew. Chem.* **2013**, *125*, 4364; b) D. Paunescu, M. Puddu, J. O. B. Soellner, P. R. Stoessel, R. N. Grass, *Nat. Protoc.* **2013**, *8*, 2440.
[15] M. E. Allentoft et al., *Proc. R. Soc. London Ser. B* **2012**, *279*, 4724.
[16] T. Lindahl, *Nature* **1993**, *362*, 709.
[17] D. Paunescu, C. A. Mora, M. Puddu, F. Krumeich, R. N. Grass, *J. Mater. Chem. B* **2014**, *2*, 8504.
[18] P. F. Campos, O. E. Craig, G. Turner-Walker, E. Peacock, E. Willerslev, M. T. P. Gilbert, *Ann. Anat.* **2012**, *194*, 7.
[19] M. Salamon, N. Tuross, B. Arensburg, S. Weiner, *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 13783.
[20] C. I. Smith, A. T. Chamberlain, M. S. Riley, C. Stringer, M. J. Collins, *J. Hum. Evol.* **2003**, *45*, 203.

