Langzeitdatenspeicher

DOI: 10.1002/ange.201411378

Robuste chemische Speicherung von digitalen Informationen auf DNA in Silicat unter Verwendung fehlerkorrigierender Codes**

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

Abstract: *Auf Papier gedruckte oder auf Mikrofilm projizierte Information kann mehr als 500 Jahre überdauern. Die Speicherung von digitaler Information über Zeiträume über 50 Jahre stellt hingegen eine große Herausforderung dar. Hier zeigen wir, dass digitale Information auf DNA gespeichert werden kann und auch nach wesentlich längeren Zeiträumen wieder fehlerfrei auslesbar ist. Um eine Wiederherstellung der gespeicherten Information zu gewährleisten, wurde DNA in eine anorganische Matrix eingeschlossen. Zusätzlich dazu wurden fehlerkorrigierende Codes verwendet, um während der Lagerung entstehende Fehler zu beheben. Hierfür codierten wir 83 Kilobytes an Information in 4991 DNA-Segmente, die jeweils 158 Nukleotide lang waren und in einer SiO_2-Matrix eingeschlossen wurden. Beschleunigte Alterungsprozesse wurden simuliert, um die Zerfallskinetik der DNA zu untersuchen. Es zeigte sich, dass die Daten in DNA unter verschiedensten Bedingungen Jahrhunderte lang archiviert werden können. Die ursprüngliche Information konnte selbst nach einwöchiger Lagerung bei 70 °C noch fehlerfrei wiederhergestellt werden. Dies ist thermisch äquivalent zu einer Lagerung in Zentraleuropa über einem Zeitraum von ca. 2000 Jahren.*

Prähistorische Informationen in der Form von Höhlenmalereien, alte Inschriften, Gravuren in Gold oder mittelalterliche Texte stellen wichtige Zeugnisse unserer Vergangenheit dar. Ein Beispiel ist das Palimpsest des Archimedes aus dem zehnten Jahrhundert, das die einzig bekannte Kopie der „Methodenlehre“ beinhaltet und einen Meilenstein in der Entwicklung der Geometrie und modernen Algebra darstellt. Dieses Buch überdauerte mehr als 1000 Jahre und wurde 1998 auf einen Wert von mehr als zwei Millionen USD geschätzt. Angesichts dieses Wertes scheint es eher überraschend, dass derzeitig nur wenige Projekte zur Entwicklung von langlebigen Speichern für digitale Daten vorhanden sind (z. B. M-Disc, Syylex). Des Weiteren ist die Halbwertszeit von Informationen seit der Umstellung von analogen auf digitale Speichermedien drastisch gesunken.[1]

Herkömmliche Speichermedien, wie z. B. optische oder magnetische Datenträger sind für die Langzeitspeicherung von Daten (> 50 Jahre) nicht zuverlässig genug.[2] Zudem benötigt die Entwicklung besserer Datenträger lange Testphasen, die die aktuellen Produktentwicklungszyklen bei weitem überschreiten. DNA ist das einzige Speichermedium, für das echte Langzeitdaten aus der Archäologie zur Verfügung stehen. Kürzlich wurde jeweils 300 000 Jahre alte mitochondriale DNA von Bären und Menschen sequenziert.[3] Zudem findet DNA Verwendung als Codierungssprache in den Bereichen der Forensik,[4] Produktmarkierung[5] und im DNA-Computing.[6] Folglich wurden bereits mehrere Methoden entwickelt, um Informationen in DNA zu speichern.[4] Bisherige Herangehensweisen sind jedoch nicht zuverlässig, da sie mit entstehenden Fehlern nicht umgehen können. Des Weiteren schlagen sie keine (physikalische) Lösung zur Lagerung von DNA vor, um die Stabilität der Informationen über viele Jahre zu gewährleisten.

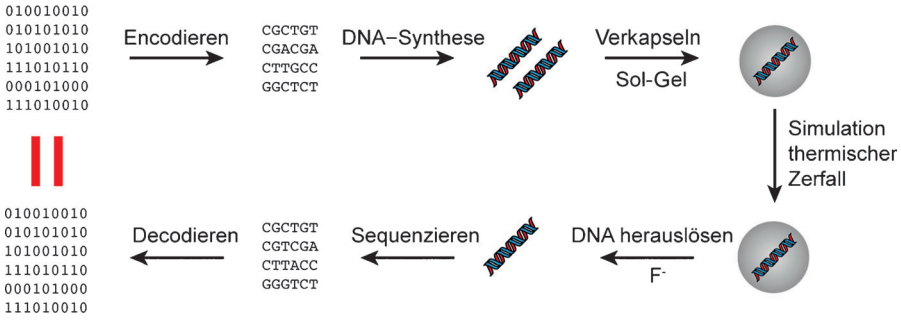
Zur Bewältigung dieser Probleme kombinierten wir ein fehlerkorrigierendes Informations-Codierungs-Schema mit einer bereits etablierten Methode zur Lagerung von DNA in „synthetischen Fossilien“ (Schema 1). Die dazugehörigen Experimente zeigen, dass nur durch die Kombination der beiden Konzepte Informationen selbst nach einer Millionen Jahre langer Lagerung in der „Svalbard Global Seed Vault“ (bei −18 °C) vollständig wiederhergestellt werden könnten.

Da die Synthese und Sequenzierung von langen DNA-Strängen technisch nur schwer möglich ist, werden die Daten auf viele kurze Segmente geschrieben. Diese können nicht geometrisch angeordnet werden, wodurch sich das Schreiben und Lesen von Daten auf DNA von herkömmlichen Speichermedien wie z. B. Festplatten unterscheidet. Zudem treten beim Schreiben, Lagern und Ablesen (Sequenzieren) der DNA Fehler auf. Einzelne Basen sind fehlerhaft, zudem gehen ganze Sequenzen verloren. In klassischen Speichermedien werden fehlerkorrigierende Codes verwendet, mit denen die Information durch das Hinzufügen von Redundanz geschützt wird. Die Redundanz wird so gewählt, dass alle Fehler, die während der Benutzung oder Lagerung der Daten auftreten, korrigiert werden können. Aufgrund der spezifischen Anforderungen an die Lagerung von DNA mussten bereits existierende Algorithmen entsprechend modifiziert und angepasst werden: Einzelne Sequenzen wurden mit Indizes versehen und mit zwei unabhängigen fehlerkorrigie-

[*] Dr. R. N. Grass, M. Sc. M. Puddu, M. Sc. D. Paunescu, Prof. W. J. Stark
Institut für Chemie- und Bio-Ingenieurwissenschaften, ETH Zürich
Vladimir-Prelog-Weg 1, 8093 Zürich (Schweiz)
www.fml.ethz.ch
E-Mail: robert.grass@chem.ethz.ch
Dr. R. Heckel
Department Informationstechnologie und Elektrotechnik
ETH Zürich
Sternwartstrasse 8, 8092 Zurich (Switzerland)

[**] Wir danken dem Institut für Chemie- und Bio-Ingenieurwissenschaften der ETH Zürich, dem Schweizerischen Nationalfonds (200021-150179) und dem EU-ITN-Netzwerk Mag(net)icFun (PITN-GA-2012-290248) für die finanzielle Unterstützung sowie Christof Wunderli (Microsynth AG) und Marcello Caraballo (Customarray Inc.) für Hilfe bei der DNA-Synthese und Sequenzierung.

Hintergrundinformationen zu diesem Beitrag sind im WWW unter http://dx.doi.org/10.1002/ange.201411378 zu finden.



Schema 1. Digitale Information wird in DNA codiert und in sphärischen Siliciumdioxidmatrizen verkapselt. Die DNA wird mittels Fluoridchemie aus der Matrix freigesetzt und durch Illumina-Sequenzierung decodiert. Damit wird die originale Information wiederhergestellt, selbst wenn während des Prozesses Fehler eingeführt wurden.

renden (Reed-Solomon) Codes verknüpft. (Abbildung 1; siehe Hintergrundinformationen für Code-Design und Parameterwahl).

Der entwickelte Algorithmus wurde physikalisch getestet, indem wir den Text aus zwei alten Dokumenten auf DNA speicherten: Der Schweizer Bundesbrief von 1291 und die englische Übersetzung von „Archimedes' Methodenlehre von Mechanischen Sätzen". Der vollständige (nicht komprimierte) Text besitzt eine Größe von 83 Kilobytes und wurde wie in Abbildung 1 gezeigt codiert. Dadurch entstanden 4991 Sequenzen zu je 117 Nukleotiden, an die zusätzlich Primer angebracht wurden (Totallänge von 158 nt), um eine schnelle und indizierte Vorbereitung der Sequenzdatenbank für die Sequenzierung zu ermöglichen. Die einzelnen Sequenzen wurden mithilfe von Mikroarray-Technologie (CustomArray)[6] synthetisiert, anschließend durch eine benutzerdefinierte PCR-Methode (Polymerasekettenreaktion) für das Sequenzieren vorbereitet und unter Verwendung der Illumina-MiSeq-Plattform abgelesen (siehe Hintergrundinformationen für experimentelle Details). Beim Ablesen der Sequenzen musste der innere Code durchschnittlich 0.7 Fehler pro Sequenz ausgleichen.

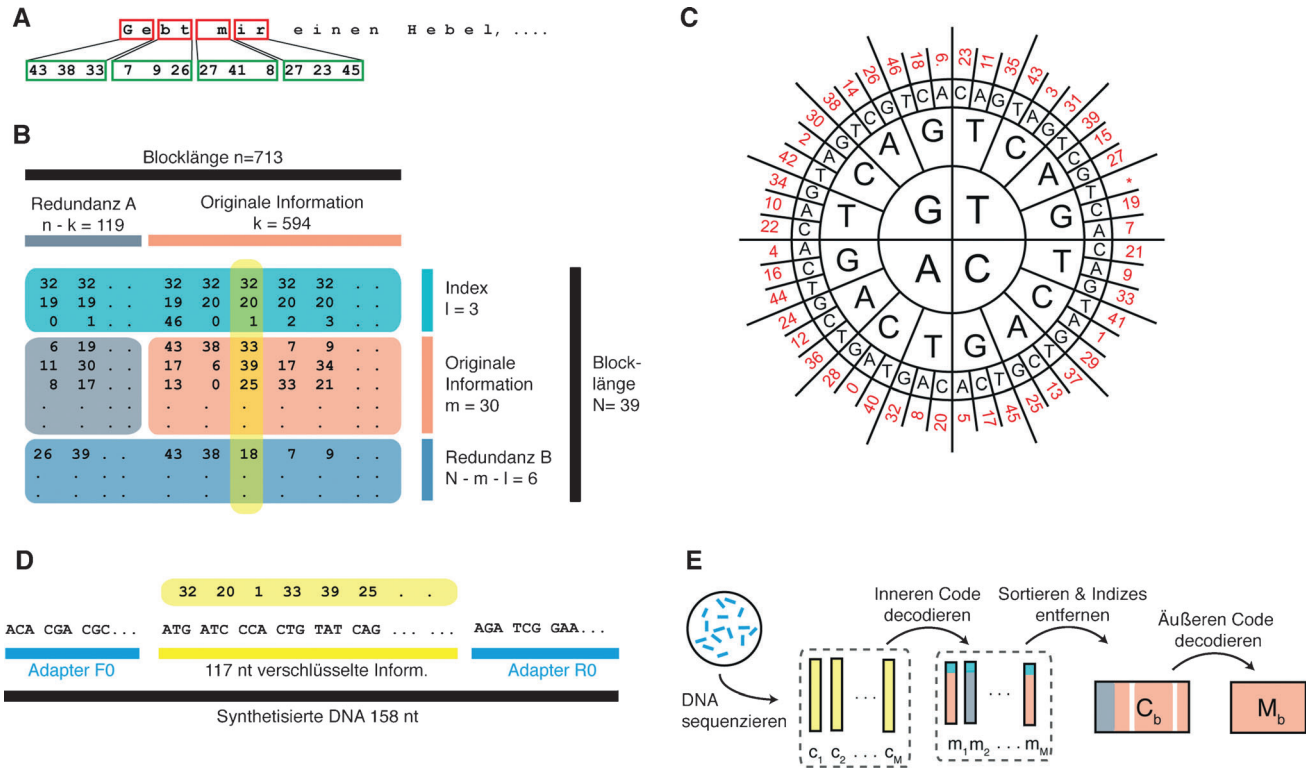


Abbildung 1. Verschlüsselung des Textes in DNA durch Reed-Solomon(RS)-Codes: A) Jeweils zwei Buchstaben (oder generell zwei Bytes eines digitalen Files) werden drei Elementen des Galois-Feldes der Größe 47, GF(47), durch einen Zahlbasiswechsel (256^2 nach 47^3) zugeordnet. Diese originale Information wird in Blöcken von 594×39 Elementen angeordnet. B) Die einzelnen Blöcke werden in einem äußeren Schritt durch einen RS-Code verschlüsselt; dadurch wird die Redundanz A hinzufügt. Anschließend wird zu jeder Spalte ein Index hinzugefügt, und in einem inneren RS-Codierungsschritt wird die Redundanz B hinzugefügt. C) Die einzelnen Spalten werden in DNA übersetzt, indem jedem Element von GF(47) drei Basen zugeordnet werden. Die Zuordnung erfolgt mit der GF(47)toDNA-Codesonne, welche gewährleistet, dass keine Base mehr als dreimal wiederholt wird. D) Zwei konstante Adapter werden hinzugefügt, und die resultierende 158 Nukleotide lange Sequenz kann synthetisiert werden. E) Um die originale Information zurückzugewinnen, wird die gelesene DNA-Sequenz zu GF(47) zurückübersetzt. Zuerst wird der innere Code decodiert, welcher die Fähigkeit hat, einzelne Fehler in der Basensequenz zu korrigieren. Dann werden die Sequenzen nach dem Index sortiert und schließlich vom äußeren Decodierer decodiert. Der äußere Decodierer kann ganze Sequenzen ersetzen und korrigieren. (Siehe Hintergrundinformationen für Details bezüglich Codierung und experimentelle Details.)

Der äußere Code musste zusätzlich einen Verlust von 0.3% der gesamten Sequenzen kompensieren und 0.4% der Sequenzen korrigieren. Dies ermöglichte eine vollständige und fehlerfreie Rückgewinnung der ursprünglich gespeicherten Information.

Dieses Experiment zeigt, dass digitale Information zuverlässig auf DNA gespeichert werden kann. Unser nächstes Experiment sollte zeigen, dass DNA tatsächlich für extrem lange Zeiträume als Speicher verwendet werden kann. Dies ist nicht selbstverständlich, da DNA in Lösung innerhalb einiger Jahre zerfällt.[10] Um herauszufinden, ob gelagerte DNA im festen Zustand stabiler ist,[11] testeten wir die Stabilität des Oligo-Pools mit 4991 Elementen anhand von drei zuvor etablierten Trockenlagerungstechnologien in beschleunigten Alterungstests (Abbildung 2). Die Lagerungstechnologien bestehen aus einem imprägnierten Filterpapier,[9] einem Biopolymer, das den glasartigen Zustand der DNA in Samen und Sporen nachahmt,[10] und einem „synthetischen Silicat-Fossil", das auf einer in unserer Gruppe entwickelten Methode basiert.[11] Verglichen mit der Lagerung von trockener DNA ohne zusätzliches Hilfsmittel verlängern alle drei Methoden die Haltbarkeit der DNA beträchtlich. Durch die Temperaturabhängigkeit der Zerfallsgeschwindigkeit und unter der Annahme einer Zerfallskinetik erster Ordnung konnten Arrhenius-Aktivierungsenergien (E_a) berechnet werden, die für alle drei Lagerungsmethoden annähernd gleich sind ($155 \pm 2\ \mathrm{kJ\,mol^{-1}}$; siehe Hintergrundinformationen für Details). Die Aktivierungsenergie liegt dabei im Bereich der vor kurzem publizierten Einzelstrangbruch-Kinetik von DNA in trockener Lagerung ($E_a = 158\ \mathrm{kJ\,mol^{-1}}$)[8] und unterscheidet sich beträchtlich von der bisher bekannten Zersetzungskinetik von DNA in Lösung (105–120 $\mathrm{kJ\,mol^{-1}}$).[7] Obwohl die Aktivierungsenergien für die drei Aufbewahrungsmethoden nahezu identisch sind, unterscheiden sich die einzelnen Zerfallsgeschwindigkeiten deutlich. Die Zerfallskinetik von getrockneter DNA in Abhängigkeit von der Luftfeuchtigkeit kann folgendermaßen ausgedrückt werden:

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

wobei der beobachtete Faktor A den präexponentiellen Arrhenius-Faktor k_0 und den Effekt des Wassers $\left(c_{\mathrm{H_2O}}\right)^n$ beschreibt. Basierend auf den identischen Aktivierungsenergien könnte man daraus schließen, dass sich die DNA in allen Lagerungsmatrizen nach demselben Einzelstrangbruchmechanismus zersetzt[11,16] und sich die einzelnen Zerfallsgeschwindigkeiten nur aufgrund der Lagerungstemperatur und der Wasserkonzentration in der unmittelbaren Nähe der DNA-Moleküle unterscheiden. (Es lässt sich vermuten, dass sich Wasser an die DNA-Moleküle innerhalb der Biopolymere anlagert, und selbst wenn DNA in Siliciumdioxid eingeschlossen ist, wird sie noch mit Wassermolekülen assoziiert sein.) Die Daten in Abbildung 2 zeigen deutlich, dass die

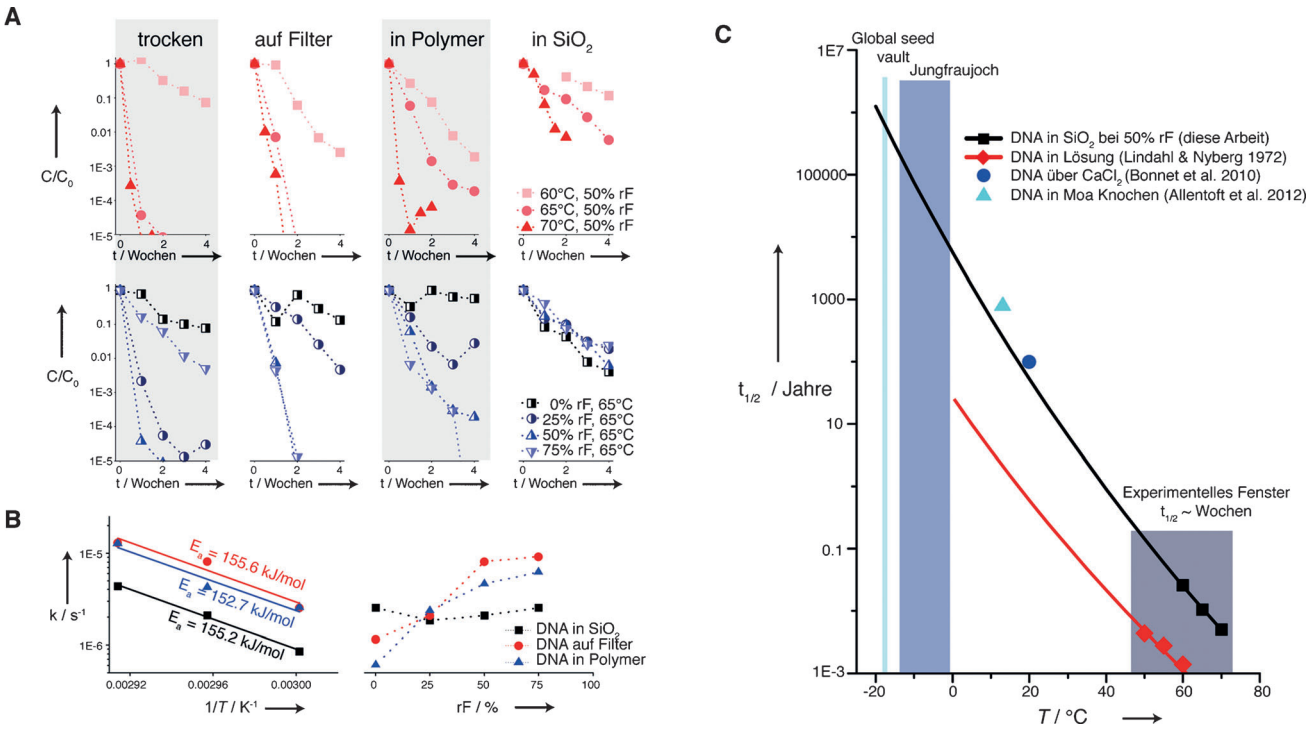


Abbildung 2. Zersetzungskinetik von trocken gelagerter DNA: A) Auswirkungen von Temperatur und Luftfeuchtigkeit auf die DNA-Konzentration (per qPCR) als Funktion der Zeit, unter Verwendung von vier verschiedenen Lagerungsmethoden: reine DNA im festen Zustand, DNA auf FTA-Filterpapier, DNA in Biopolymer (DNAstable) und DNA eingeschlossen in Siliciumdioxid (Messungen bei 70°C wurden aufgrund der beschleunigten Kinetik häufiger aufgenommen). B) Davon abgeleitete Geschwindigkeitskonstanten erster Ordnung und entsprechende Aktivierungsenergien. C) Die Halbwertszeit von DNA eingeschlossen in Siliciumdioxid wurde mit einer Arrhenius-Aktivierungsenergie von $155 \pm 10\ \mathrm{kJ\,mol^{-1}}$ extrapoliert und mit Literaturdaten zur DNA-Stabilität in Lösung,[7] getrockneter DNA[8] und DNA in fossilen Knochen verglichen.[12] Die Literaturwerte sind skaliert auf:[12] $t_{1/2}{}^{158\,\mathrm{nt}} = t_{1/2}{}^{1\,\mathrm{nt}}/158$.

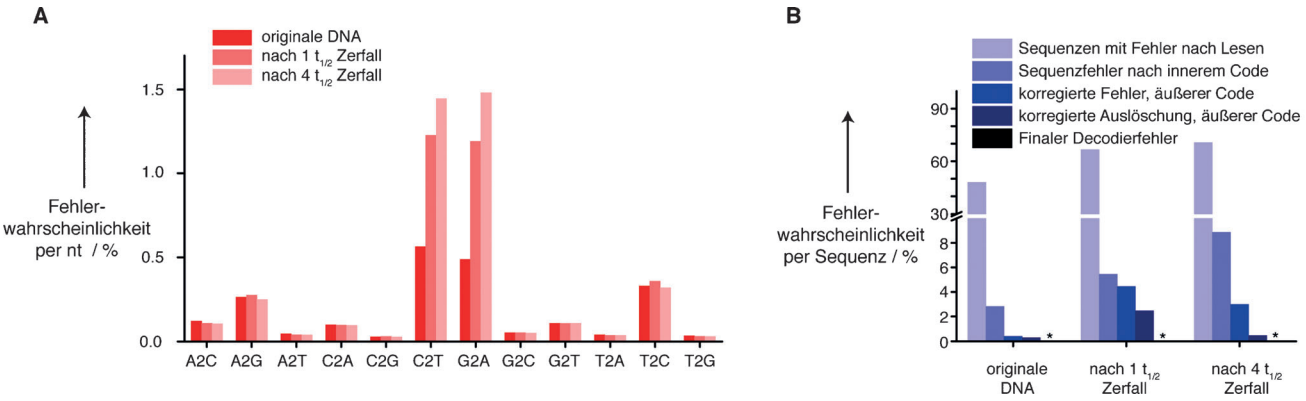


Abbildung 3. Statistik zur Entschlüsselung und Fehlerkorrektur: A) Fehlerwahrscheinlichkeit beim Ablesen einzelner Basen (beispielsweise bezeichnet A2C die Wahrscheinlichkeit, dass die Base C anstatt A gelesen wird). B) Wahrscheinlichkeit von Fehlern in einer Sequenz (Fehler des inneren Codes) sowie Sequenzfehler und Auslöschungen, die allesamt durch den inneren und äußeren Algorithmus behoben werden. Somit ergibt sich eine gesamte Fehlerwahrscheinlichkeit von 0 (*) für alle drei Fälle. (Siehe Hintergrundinformationen für Definitionen und nähere Beschreibung der Fehler.)

anorganische Lagerungsmethode (DNA in Siliciumdioxid) die beste DNA-Konservierungsmethode darstellt, da sie die geringste lokale Wasserkonzentration aufweist. Eine anorganische Schicht trennt zudem die DNA-Moleküle von der Umgebung, und dadurch wird der Zerfall nicht von der Luftfeuchtigkeit der Lagerungsumgebung beeinflusst. Für eine Langzeitlagerung ist dies von großer Bedeutung, da Lagerstätten ohne Feuchtigkeit schwer aufrechtzuerhalten sind. Im Gegensatz dazu können alterungshemmende Faktoren wie Kälte (beispielsweise Permafrost) und der Ausschluss von Licht für längere Zeit ohne Energieaufwand verwirklicht werden. Das Lagerungsverfahren in Siliciumdioxid weist außerdem einen außergewöhnlichen Schutz der DNA gegenüber Oxidation auf (siehe Behandlung mit reaktiven Sauerstoffspezies in den Hintergrundinformationen). Ein weiterer Schutz gegenüber Licht kann durch den Einsatz einer Titandioxid-Schicht erzielt werden.[17]

In fossilen Knochen hat DNA die größte Überlebenschance, wenn sie in Apatit/Kollagen-Strukturen[13] oder in Kristallaggregaten[14] eingeschlossen ist. Diese Strukturen schützen die DNA vor der Umwelt und Feuchtigkeit, ähnlich zur Einkapselung der DNA im anorganischen Siliciumdioxid, welche wir hier verwendet haben. Werden die in Abbildung 2 gezeigten Zerfallsdaten der DNA für tiefere Temperaturen extrapoliert, stimmen sie sehr gut mit der Zerfallsgeschwindigkeit von Moa-Fossilien überein, welche Allentoft et al. anhand von bis zu 8000 Jahren alten Knochen untersucht haben.[12] Des Weitern stimmen die Daten auch mit den kürzlich bestimmten Zerfallsdaten von trocken gelagerter DNA überein (32 Wochen Lagerung; Punkt 4 in Bonnet et al.[8]). Diese Stabilität erklärt zudem den Erfolg der Sequenzierung von DNA aus 300000 Jahre alten Knochenproben (siehe Diskussion in den Hintergrundinformationen). Daraus wird ersichtlich, dass in beiden Fällen (DNA in Knochen und DNA in Siliciumdioxid) der Zerfall an Informationen derselben Kinetik folgt. Die beschleunigten Alterungstests von DNA in Siliciumdioxid ahmen dabei den langzeitigen Zerfall von DNA in fossilen Knochen nach.

Im Folgenden wollen wir zeigen, dass in synthetischer DNA gespeicherte Information selbst nach beträchtlicher thermischer Behandlung immer noch korrekt ausgelesen werden kann. Hierfür wurde DNA in Siliciumdioxid für eine halbe sowie für eine ganze Woche bei 70 °C gelagert und anschließend sequenziert. Die Daten wurden dann nach dem vorherigen Schema (Abbildung 1) rekonstruiert. Die zwei erhaltenen Datenpunkte entsprechen dem Zerfall der DNA zu ca. einer bzw. vier Halbwertszeiten. Auch wenn in diesen thermisch behandelten Proben beide (innere und äußere) fehlerkorrigierende Codes signifikant mehr Fehler korrigieren mussten als in den thermisch unbehandelten Proben, konnte in beiden Fällen die Information ohne Fehler wiederhergestellt werden (Abbildung 3).

Die Möglichkeit, Ursprungsdaten aus DNA noch nach 4 Halbwertszeiten fehlerfrei auslesen zu können, entspricht nach Abbildung 2c der Lagerung von DNA in Zürich (9.4 °C) für 2000 Jahre oder für den am kältesten, ganzjährlich zugänglichen Punkt in der Schweiz (Jungfraujoch, 3471 m.ü.M) bis zu 100000 Jahre. Die Daten sagen zudem voraus, dass digitale Information in eingekapselter DNA in Siliciumdioxid für mehr als 2 Millionen Jahre im Global Seed Vault bei −18 °C gespeichert werden kann.

Eingegangen am 24. November 2014
Online veröffentlicht am 30. Januar 2015

Stichwörter: DNA · Fossilien · Informationsspeicherung · Langzeitdatenspeicher · Sol-Gel-Prozesse

[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.