A Bountiful Harvest of Rainwater

OVER THOUSANDS OF YEARS, SOCIETIES HAVE developed a diversity of local water harvesting and management regimes that continue to survive in South Asia, Africa, and other parts of the world (*1*). Such systems are often integrated with agroforestry (*2*) and local forest management practices (*3*). In their Policy Forum "Managing water for people and nature" (*Science*'s Compass, 11 May, p. 1071), Nels Johnson and co-authors discuss several market mechanisms for sustainable water management, including taxing users to pay commensurate costs of supply and distribution and costs of integrated watershed management, and charging polluters for effluent treatment. Although such measures are indeed essential, I would argue that they are insufficient: They should be complemented with policy innovations to promote rainwater harvesting (*4*).

Revival of local practices of rainwater harvesting could provide substantial amounts of water. For example, a hectare of land in Barmer, one of India's driest places, with 100 millimeters of rainfall annually, could yield 1 million liters of water per year from harvesting rainwater. Even with simple technology such as ponds and earthen embankments called tanks, at least half a million liters a year can be harvested from rain falling over 1 hectare of land, as is being done in the Thar Desert, making it the most densely populated desert in the world. Indeed, there are 1.5 million village tanks in use and sustaining everyday life in the 660,000 villages in India.

In the Negev Desert, decentralized harvesting of water in microcatchments from rain falling over a 1-hectare watershed yielded 95,000 liters of water per hectare per year, whereas collection efforts from a single large unit from a 345-hectare watershed yielded only 24,000 liters per hectare per year (*5*). Thus, 75% of the collectible water was lost as a result of the longer distance of runoff. Indeed, this is consistent with local knowledge distilled in an Indian proverb, "Capture rain where it rains."

In the cities, rainwater could be harvested from building rooftops for residential use, and any surplus could be channeled through bore wells to replenish the groundwater, avoiding loss to runoff. However, if tanks and other rain harvesting technology are to be used to their full potential, policy innovations must include institutional changes so that such common-pool resources are effectively managed (*6*). Also, all forms of government subsidies need to be removed to allow market mechanisms, such as the ones Johnson *et al.* discuss, to run their course. Users would then find it prudent not only to make efficient use of priced water, but they would also have the incentive to collect the gift that Mother Nature has to offer in the form of rain.

Villages in the Thar Desert in India, the world's most densely populated desert, rely heavily on the harvesting of rainwater for their daily needs.

DEEP NARAYAN PANDEY
Indian Forest Service, Indian Institute of Forest Management, Bhopal 462003, India. E-mail: dnpandey@vsnl.com or deep@inef.org

References and Notes
1. A. Agarwal, S. Narain, Eds., *Dying Wisdom: Rise, Fall and Potential of India's Traditional Harvesting Systems* (Centre For Science and Environment, New Delhi, 1997).
2. H. R. Wagachchi, K. F. Wiersum, *Agrofor. Sys.* **35**, 291 (1997).
3. D. N. Pandey, *Ethnoforestry: Local Knowledge for Sustainable Forestry and Livelihood Security* (Himanshu/AFN, New Delhi, 1998).
4. T. M. Boers, J. Ben-Asher, *Agric. Water Manag.* **5**, 145 (1982).
5. M. Evenari, L. Shanan, N. Tadmor, *The Negev: The Challenge of a Desert* (Harvard Univ. Press, Cambridge, ed. 2, 1982).
6. E. Ostram *et al.*, *Science* **284**, 278 (1999).

Letters to the Editor

Letters (~300 words) discuss material published in *Science* in the previous 6 months or issues of general interest. They can be submitted by e-mail (science_letters@aaas.org), the Web (www.letter2science.org), or regular mail (1200 New York Ave., NW, Washington, DC 20005, USA). Letters are not acknowledged upon receipt, nor are authors generally consulted before publication. Whether published in full or in part, letters are subject to editing for clarity and space.

Long-Term Storage of Information in DNA

IN THIS DIGITAL AGE, THE TECHNOLOGY USED for information storage is undergoing rapid advances. Data currently being stored in magnetic or optical media will probably become unrecoverable within a century or less, through the combined effects of hardware and software obsolescence and decay of the storage medium. New approaches are required that will permit retrieval of information stored for centuries or even millennia.

DNA has three properties that recommend it as a vehicle for long-term information storage. First, DNA has stood the informational "test of time" during the billions of years since life emerged. Nonreplicating DNA molecules are also quite robust. Although DNA stored under nonideal conditions (e.g., in archaeological deposits) is subject to hydrolytic and oxidative damage (*1*), mitochondrial DNA extracted and amplified from 7000-year-old human remains yielded an accurate DNA sequence (*2*). Storage of DNA under more favorable conditions can result in extremely long stability, as evidenced by the reported recovery of viable bacteria from 250-million-year-old salt crystals (*3*). Second, because DNA is our genetic material, methods for both storage and reading of DNA-encoded information should remain central to technological civilizations and undergo continual improvements. Third, use of DNA as a storage medium would permit each segment of information to be stored in an enormous number of identical molecules. This extensive informational redundancy would strongly mitigate effects of any losses due to stochastic decay (*4*).

Data retrieval of information stored in DNA should ideally require minimal prior knowledge beyond a familiarity with molecular biological techniques. In the procedure we have developed, two standard techniques are required for recovery of stored information: polymerase chain reaction (PCR) and DNA sequence analysis. Central to our pro-

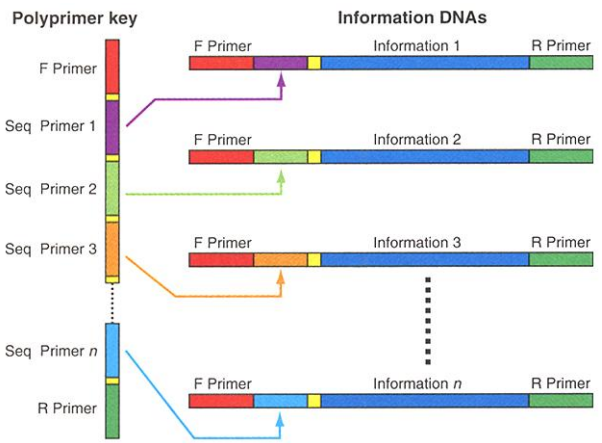
cedure is the use of two classes of DNA (see the figure): information DNAs (iDNAs) containing the stored information, and a single polyprimer key (PPK) that is the key to retrieving the information stored in the iDNAs. Each iDNA contains the following sequence elements: common flanking forward (F) and reverse (R) PCR amplification primers (~10 to 20 bases long), a unique sequencing primer (comparable in size to the F and R primers), a small common spacer (~3 to 4 bases long) serving as a cue to indicate the start of the stored information, and a unique information segment. The information to be stored is encoded successively in these information segments, beginning with Information 1. The PPK is also flanked by the common F and R primers and contains in the proper order the unique sequencing primers for the ordered retrieval from each iDNA of its information segment sequence. Common spacer sequences indicate the demarcations between each sequencing primer.

Each information segment should be capable of encoding any possible data (e.g., text), whereas correct readout requires that each sequencing primer prime a sequencing reaction only from the appropriate position within a specific iDNA, and not misprime on any iDNA. Various approaches can be taken to satisfy these conditions (*5*).

To retrieve the stored information, a future reader would proceed as follows. First, the PPK is amplified and sequenced, by using co-stored F and R primers. PCR amplification would yield amounts of the PPK sufficient for further analysis, even if extensive degradation or modification had occurred during storage. Sequence analysis of the entire PPK would reveal the sequences of the F and R PCR primers, plus an internal ordered series of elements of comparable size, suggesting roles for these elements as sequencing primers. This interpretation would lead the reader to the second step, sequence analysis of the information segments. Assuming that the F and R primers were interpreted to be "universal" PCR primers, the reader would use these for simultaneous PCR amplification of all of the collectively stored iDNAs. Sequential use of each sequencing primer to prime a sequencing reaction on the collection of PCR products would then yield the sequence of each of the information segments, arranged in the proper order to be decoded and read as a continuous block.

We have carried out a simple prototype of this technique, using only the bases A, C, and T to encode English text based on an "obvious" ternary code (*5, 6*). However, even if this encoding were not obvious to a future reader, recovery of a sufficient number of information segment sequences would permit use of standard cryptanalytical techniques to determine how text had been encoded in the DNA.

Two iDNAs were constructed to encode, respectively, "IT WAS THE BEST OF TIMES IT WAS THE WORST OF TIMES," and "IT WAS THE AGE OF FOOLISHNESS IT WAS THE EPOCH OF BELIEF." Not only is this text one of the most famous opening lines of a novel (*7*), but the fourfold repetition of the phrase "it was the" provided a test of the ability of this approach to deal with repeated DNA sequences both within and between information segments. We simultaneously amplified by PCR the two iDNAs, and then sequenced each of the products; decoding of the resultant DNA sequences successfully recovered the stored text (*8*).



Structures of DNA molecules used for information storage and readout. The single polyprimer key contains a series of sequencing primer sequences (Seq Primer #) flanked by common forward (F Primer) and reverse (R Primer) PCR primer sequences, each separated by a small common spacer. Each information DNA is also flanked by the common PCR primer sequences and contains two unique elements, separated by the common spacer: a Seq Primer and a numbered information segment (Information #) (not drawn to scale).

The two standard molecular biological techniques used in our model for information storage in DNA could form the basis for a variety of DNA-based memory storage devices (*9*). Moreover, the use of DNA microarray technology should permit extensive scale-up of this model. Current microarray technology, in which up to 10,000 small DNA samples can be spotted in an ordered array onto an approximately 3-square-centimeter surface (glass, etc.) (*10*), would have to be modified to permit spotting of DNA into small wells at a comparable density on a "microchip." Use of such a microchip for storage would impose two levels of order on the information stored in iDNAs placed in these microwells: the x and y coordinates of each microwell, plus the order within each microwell provided by the scheme described here. A single series of unique identification primers, encoded within a single PPK, should suffice to order the collection of iDNAs within every microwell and thus permit readout in the proper order of the information stored on the entire chip. Because of the enormous number of different potential 20-base primer sequences (*11*), the capacity for information storage in microarrayed DNA is presently limited by practical rather than theoretical considerations. It seems reasonable that with minor advances in microarray technology, about 200 novels or other data each equivalent in size to *A Tale of Two Cities* could be stored in a DNA microchip with the area of a postage stamp (*12*). Ongoing technological advances should greatly increase this capacity.

CARTER BANCROFT,[1]* TIMOTHY BOWLER,[2] BRIAN BLOOM,[1] CATHERINE TAYLOR CLELLAND[1]
Departments of [1]Physiology and Biophysics, and [2]Biochemistry and Molecular Biology, Mount Sinai School of Medicine, New York, NY 10029, USA
*To whom correspondence should be addressed. E-mail: carter.bancroft@mssm.edu

References and Notes

1. M. Hoss, P. Jaruga, T. H. Zastawny, M. Dizdaroglu, S. Paabo, *Nucleic Acids. Res.* **24**, 1304 (1996).
2. S. Paabo, J. A. Gifford, A. C. Wilson, *Nucleic Acids Res.* **16**, 9775 (1988).
3. R. H. Vreeland, W. D. Rosenzweig, D. W. Powers, *Nature* **407**, 897 (2000).
4. Studies of ancient human remains provide information on long-term rates of DNA decay and/or modification, and thus a highly conservative estimate of minimum DNA amounts required for prolonged storage under more ideal conditions. About 0.1% of the DNA extracted from ancient, decomposed tissue is unmodified [(*1*); S. Paabo *et al.*, *J. Biol. Chem.* **264**, 9709 (1989)]. However, as little as 100 to 300 femtograms of this unmodified DNA can serve as a PCR template (*2*). In the prototype we have executed, information is stored in 20 nanograms (20,000 picograms) of identical DNA molecules ~250 base pairs in size [see text and (*8*)], far above the 100-picograms range. Moreover, use of this large number of molecules (about 80 billion) as PCR templates in our readout procedure should greatly suppress effects of any base modifications during prolonged storage that yield sequence changes in individual DNA molecules [O. Handt *et al.*, *Am. J. Hum. Genet.* **59**, 368 (1996)].
5. In the simplest model, two bases (e.g., A and T) would be used to encode text in information segments, and the other two (e.g., G and C) to construct sequencing primers. This would prevent mishybridization of sequencing primers to information segments, but would greatly limit efficiency of text storage and the number of possible different sequencing primers. In the prototype we have executed, we have instead encoded text using only the bases A, C, and T. Sequencing primers were designed with all four bases, plus a requirement that each fourth position be a G. The resultant mismatch at (at least) each fourth position between the sequences of any sequencing primer and any information segment should prevent mispriming. Scale-up of the storage model presented here would ultimately require computer-generated design [see, for example, M. Garzon *et al.*, in *DNA Based Computers V*, volume 54 of *DIMACS Series in Discrete Mathe-*

matics and Theoretical Computer Science, E. Winfree, D. K. Gifford, Eds. (American Mathematical Society, Providence, RI, 2000), p. 91] of large numbers of both sequencing primers and iDNA sequences that satisfy the constraints on these elements.
6. The DNA bases were ordered alphabetically (A, C, T). DNA codons were then constructed by means of a ternary code, beginning with "AAA" (encoding the letter "A"). The bases C, and then T, were inserted progressively into the third, second, and first positions, yielding a series of 27 codons encoding the English letters in alphabetical order, plus a space (*8*).
7. C. Dickens, *A Tale of Two Cities* [Oxford Univ. Press, London, New York, 1953 (originally published in 1859)].
8. Supplementary material is available at http://www.sciencemag.org/cgi/content/full/293/5536/1763/DC1
9. The combined operations of PCR followed by sequence analysis are analogous to the retrieval of information from an addressable storage device such as the random access memory in a computer. The ability to use these combined operations to retrieve data permits construction of DNA representations of classical computer data structures such as arrays, linked lists, and trees. The model depicted in the figure is somewhat analogous to an array data structure. The PPK contains the addresses of the data elements (the iDNA segments), which can be calculated (sequenced) and then used for selective retrieval of the stored data. In an alternative serial model, analogous to a linked list, a series of iDNAs could be designed, each containing both a data element and the sequencing primer for retrieving the data element from the succeeding iDNA. Such a serial model would obviate the need for a separate PPK, but information retrieval would require considerably more experimental manipulations (and prior specific knowledge) than in the parallel model explored here.
10. D. Gerhold, T. Rushmore, C. T. Caskey, *Trends Biochem. Sci.* **24**, 168 (1999).
11. There is a theoretical maximum of 4^{20}, but this number would be reduced by the requirement that primer sequences be designed to avoid mispriming at inappropriate sites.
12. A conservative upper limit on the size of the information segment is ~600 bases, set by the present limits on DNA sequence obtainable from a single sequencing primer. If four-base codons chosen from our three-base alphabet (A, C, T) were used (to permit encoding of all common English alphanumeric characters plus a space), each iDNA could store about 150 characters. Storage of *A Tale of Two Cities*, containing 742,901 alphanumeric characters plus spaces, would require ~5000 iDNAs. Current technology would permit single-pass sequence analysis of a single PPK containing up to 100 unique sequencing primers, implying that information could be stored in ~100 different iDNAs per microwell. Since 50 microwells would thus be required to store Dickens' novel in DNA form, a 10,000-well microchip could store ~200 texts.
13. Supported by Defense Advanced Research Projects Agency/National Science Foundation grant CCR–9724012.

The Challenge of Defining Disease

THE PHILOSOPHICAL DEFINITION OF DISEASE based on impairment or limitation of normal function that Boorse (*1*) proposed is rejected as being clinically impractical by L. K. F. Temple and co-authors in their Essay "Defining disease in the genomics era" (*Science*'s Compass, 3 Aug., p. 807). However, the definition they offer seems too broad and also fraught with its own set of difficulties.

Temple *et al.* write that "disease is *a state* that places individuals at *increased risk* of *adverse consequences*." According to this definition, activities such as mountain climbing or bungee jumping could be construed as disease states. If we could provide appropriate modifiers for the words "state" and "adverse consequences," then we would be better poised to begin a more precise definition of disease in the genomics era.

> **"...activities such as mountain climbing or bungee jumping could be construed as disease states."**