Vinton G. Cerf

DOI:10.1145/2831359

Memory with a Twist

I continue to look for ways to preserve digital information over long periods of time. Often, it is not the medium that fails but, rather, the necessary reading gear. For something

considered extremely important, it is not unusual to find a way to read only media. Think about laser scanning of long-playing records, for example. But for most users, such extreme measures may be infeasible for reasons of cost. Of course, even if the binary information can be recovered, there is still the problem of correctly decoding/interpreting the information. This is typically the job of the software that created it in the first place. I have written before about the OLIVE project at CMU that supports the execution of old operating systems and applications on computer hardware that is emulated through virtual machines.^a

In this column, however, I want to explore an idea I recently learned about in a 2013 article in *Nature* magazine.^b I spoke with two of the authors recently. The basic idea is to encode binary data in base 3, Huffman code the result, and then synthesize DNA using the standard base-pair triplets in a ternary encoding scheme. Care is taken not to use the same base pair in sequence to avoid homopolymers that might confuse the decoding stage. Additional “bits” can be used for indexing and error correction. Moreover, each fragment of about 100 bases is made to overlap with previously encoded fragments by about

75 base pairs, creating a fourfold redundancy in the encoded strings. On top of this, one can use additional base pairs, not to encode bits but to allow the sequencing equipment to select subsets of the encoded strings to avoid having to decode everything to get at only the content desired. One might think of that as a sort of biochemical file folder mechanism.

In today's world, the cost of synthesizing DNA is non-trivial but as we have seen with computing and other high technologies, costs come down and what was once laboratory-only equipment has become, literally, a household item. Setting this important cost question aside for a moment, I learned that DNA, once dehydrated, is remarkably stable. It could survive in dry condition for tens of thousands or even hundreds of thousands of years. Successful resequencing of prehistoric DNA illustrates this feature. One could imagine storing petabytes of information in vials containing dried DNA and using robots like those in today's tape libraries to retrieve an appropriate vial whose contents are to be rehydrated and sequenced. The authors report they achieved a storage density of 2.2PB/gram.

I was reminded of the IBM 1360 Photo-Digital Storage System^c that used capsules of stiff microfilm to store information. I actually had a chance to watch one of the machines that had been installed at Lawrence Livermore National Laboratory. The

electromechanical design of the system was astonishing and I found myself mesmerized as capsules of film moved from their storage locations to readers that mechanically popped open the capsules, extracted the film, read it, replaced it into the capsule, and returned the capsule to a storage bin. There could be up to 13 capsules “in flight” at one time in the system using a pneumatic tube system not unlike the way banking and payment systems once worked in department stores. A programmable controller kept track of the capsules.

While on the subject of the stability of DNA, it has occurred to me (and, of course, many others) that the theory behind *panspermia*^d is made somewhat more credible by the observation that dehydrated DNA is so stable that passage through the vacuum of space would likely not harm it. Over long periods, though, one might find the DNA is altered at the molecular level by radiation. Since travel over interstellar distances could take hundreds of thousands to millions of years, there appears to be reason to be somewhat skeptical about successful propagation of life in this fashion. For the same reason, slow interstellar propagation of information might be similarly improbable. □

d <http://www.panspermia-theory.com/>

Vinton G. Cerf is vice president and Chief Internet Evangelist at Google. He served as ACM president from 2012–2014.

Copyright held by author.

a <https://olivearchive.org/>

b N. Goldman, P. Bertone, S. Chen, C. Dessimoz, E.M. LeProust, B. Sipos and E. Birney. Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. Letter, *Nature* 494, (Feb. 7, 2013), 77 ff; doi: 10.1038/nature11875.

c https://en.wikipedia.org/wiki/IBM_1360