The New York Times

January 28, 2013

Double Helix Serves Double Duty

By JOHN MARKOFF

Last Wednesday, a group of researchers at the European Bioinformatics Institute reported in the journal Nature that they had managed to store digital information in synthetic DNA molecules, then recreated the original digital files without error.

The amount of data, 739 kilobytes all told, is hardly prodigious by today's microelectronic storage standards: all 154 of Shakespeare's sonnets, a scientific paper, a color digital photo of the researchers' laboratory, a 26-second excerpt from the Rev. Dr. Martin Luther King Jr.'s "I have a dream" speech and a software algorithm. Nor is this the first time digital information has been stored in DNA.

But the researchers said their new technique, which includes error-correction software, was a step toward a digital archival storage medium of immense scale. Their goal is a system that will safely store the equivalent of one million CDs in a gram of DNA for 10,000 years.

If the new technology proves workable, it will have arrived just in time. The lead author, the British molecular biologist Nick Goldman, said he had conceived the idea with a colleague, Ewan Birney, while the two sat in a pub pondering the digital fire hose of genetic information their institute is now receiving — and the likelihood that it would soon outpace even today's chips and disk drives, whose capacity continues to double roughly every two years, as predicted by Moore's law.

The telephone interview with Dr. Goldman, from his laboratory in Hinxton, near Cambridge, has been edited and condensed.

Does your experiment suggest that DNA is a reasonable alternative for archiving digital information?

It's too far beyond us at the moment because of the price. I don't know if there are enough machines to write DNA in big quantities. I suspect not. The experiment we did converted about three-quarters of a megabyte of information off a hard disk drive into DNA. We showed it worked on a large scale, and part of what we published is an analysis of how that might scale up, at least theoretically. But we couldn't do the scale-up experiments.

You've proved something. What's next?

We've got a couple of ideas to pursue to make this a bit more likely to be something to turn up in

the real world. One is to improve the coding and the decoding to see if we can get more information into the same amount of DNA. Hopefully if we can store twice as much information, that will halve our costs.

We were quite conservative in the approach we took. We really wanted to make sure that it worked, and so we used quite a lot of error-correction code. We could maybe sacrifice less to the error-correction part and use more actual information.

The other thing to make it work on a scale that the world would really be interested in is to automate and miniaturize. All the technologies exist — they're all commercially available. But they're not all in one place, and they're not designed to work with each other as such.

If you wanted to do it properly you'd invest in the site, you'd have DNA synthesis at the site, you'd have the storage there, you'd have the reading back in one place, and you'd miniaturize it all. You'd have micro-fluidics to do what is currently lab science — even to the level of having robots to do the filing of the test tubes onto shelves. Robots are used in magnetic tape archive centers now, and you'd just want a smaller version of the same.

How similar is what you've done to what is involved in today's gene-sequencing systems, which read and store the proteins in a DNA molecule?

The sequencing, or reading it back, that we did is exactly the same. We designed it that way. We designed it so that it would work in the standard protocols that we and our laboratory collaborators are familiar with, day in day out. It is really exactly the same process. We use an Illumina sequencing machine.

The writing of the information is a technology I'm a little bit less familiar with. But Agilent Technologies, whom we worked with, is one of the world leaders in developing this, and it is, I believe, very much like an inkjet printing system. But you're not using colored dyes on paper — you're using chemical solutions that include in them the nucleotides, the basis of DNA, fired very accurately onto a glass slide so that each little spot on the slide you build up is a separate sequence.

Is there a category of information you were most interested in archiving?

The inspiration for the project came through the issues we're having to deal with at the European Bioinfomatics Institute, where many of the authors work. We're responsible for creating and archiving and maintaining and providing to the world over the Internet some of the major biological databases: genome sequence databases, protein structure databases and others.

And we have a constant management headache. On the one hand, it's our duty to archive that information and serve it live over the Internet, but it's increasing exponentially, and as you might imagine, our budgets are not increasing exponentially. And so we have for a number of years have

had headaches, such as “Can we afford that many hard drives?” and “Can we afford to run them?” and “What are we going to do if we can’t?”

Ewan Birney, who is one of the authors of the paper, and I were joking about this in the pub. We sat down and I said: “Well, look, DNA is a really efficient way of storing information. Is there something we can do?” And as we bought another beer and got a few napkins out, we realized that on a somewhat interesting scale that we could actually do all of the component parts of something that would at least in principle scale to something that might be valuable.

One of the challenges faced in designing some organic nano-electronic components is that switches made from these molecules have been slow. Can you speed up reading and writing DNA?

The writing is increasing by a factor of 10 every five years or so. I would suspect from what people have hinted at that actually it’s going to go a bit faster than that. We’re not going to compete with silicon, I think, for speed. The main use is as a repository for high-value information that you want to keep safe, but if you really needed to go and get it you’d be prepared to wait a little while.

Have you already run out of storage space?

In some of the databases it’s gotten very close to that. They don’t just store the genomes, but they store the raw data: the output of the Illumina machine before you’ve worked out what the genome you’re studying really is. That’s part of the output of the experiment, so people would like to record that information, and we’re getting to the point where it has to be compressed in order to store it.

We’re getting to the point where we have to use “lossy” compression, so we’re beginning to lose information. There’s been a lot of discussion in that field about what can we afford to lose and how much can we afford to lose. That field is sort of on the edge of deciding what we are going to throw away. We’re absolutely outrunning Moore’s law.