Alex Berger's, Robin Reutimann's & Pascal Grumbacher's

TERM PAPER

GKG

Class 4A

Contents

BIOLOGICAL DATA STORAGE

Preface

We decided to write our term paper about the possibilities of biological data storage because we found it particularly interesting and important for the future. Very promising techniques to store digital data in synthesized DNA cells are being developed as I am writing these words. What sounds like science-fiction might be the future of our archives. Just to clarify one thing right at the beginning: Don't even think about carrying around your favourite music or movies in your own DNA. Everyday use is not the intended.

In our research documented on the following pages, we tried to find out more about this very interesting topic. What's the idea behind it? How far is the development? How could it influence the future of digital archiving? What are possible problems of the technique?

Those are some of the questions we hope to answer in this paper.

Introduction

The 21st century is the age of the computer. Over the last 15 years, information technology made huge steps and became accessible to almost everyone around the globe. Along with the development, the sheer size of all the data collected reaches unimaginable numbers:

Experts from the university of Berlin estimate the total (worldwide) amount of data collected by 2012 to be over 2.8 zettabyte (2'800'000'000'000'000'000'000 bytes)! This enormous number is expected to multiply constantly over the next few years. Sooner or later, conventional methods of storage will reach their limits. Furthermore, it isn't known how long data on a CD/ DVD or a hard drive can be conserved, the technologies are simply too young. These are the main problems which keep scientists busy all over the world, attempting to find new ways of storing large amounts of information. One of the most promising attempts is to store data in DNA. The idea itself is not completely new: Japanese scientists were thinking of using sperm for storage space in the late 90's. They estimated a single sperm cell to carry up to 780 MB of genetical information. However, the idea was given up soon, as there was no technique available to get the data into the cell, not to mention how to code or store it. In 200x, the UK scientist team around Nick Goldman picked up a similar idea:

The idea was to store information in synthesized DNA strands. This would enable huge amounts of data to be stored in a very small space. In addition synthesized DNA can be easily maintained over a long period of time. This would make the technique very suitable for infrequently accessed digital archiving. The idea is not to replace everyday-storage facilities (like hard disks in our computers, or memory cards etc.) but to find a solution for reliable long term storage. The main problems at the moment are the slowness as well as the extremely high cost of the process.. At the moment, the cost for storing only one megabyte of data hits mind-

blowing 12'400$! However, current trends in technological advances are reducing the cost of DNA synthesis at a rapid pace. The speed of the process could be massively increased by parallelization. The cost and the relatively high error rate remain to be the biggest barriers at the moment.

Engineering technique

There are different ways to process the data and each way results in different DNA. We will only describe the way that Nick Goldman used in his experiment.

Encoding

The encoding process has much more to do with mathematics and computer science than with biology. Nevertheless some steps of have a great importance for reducing errors, for example to successfully decode data from damaged DNA strands.

Every computer file is stored as a string of bytes. A byte is a binary (base-2) number of 8 digits (bits; either 0 or 1), so all numbers from 0 to 255 can be stored as a byte.

To save a file in DNA, the information firstly has to be converted from the base-2 system to the base-3 system with three instead of two possible digits (now trits; either 0, 1 or 2). Goldman used a given Huffman code for converting which also compresses the data. Afterwards, the string will be stretched with 0s, so that the number of digits are a multiple of 25.

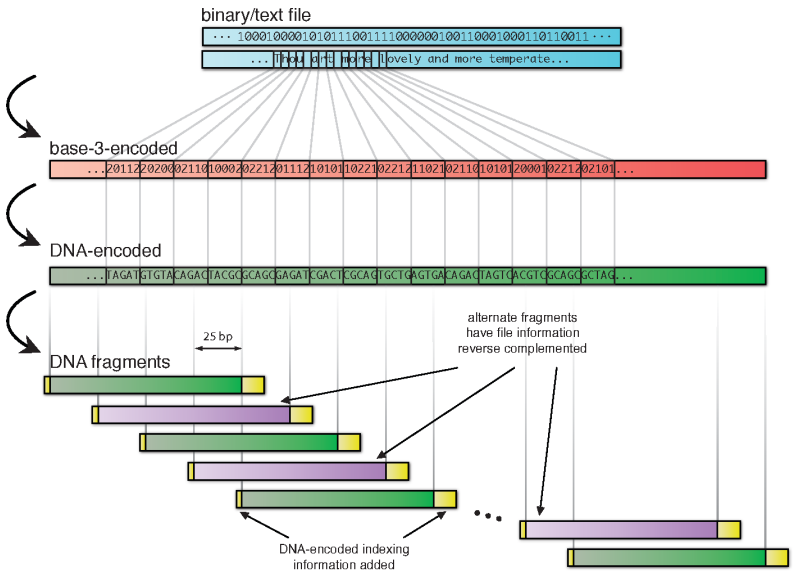
The string is now ready to be represented as DNA nucleotides. To prevent homopolymers (2 or more identical bases following each other) which led to errors in recent experiments, a trit is not simply associated with one base but the base is chosen by the previous base as the following table shows. The first trit of the string is coded using the A row.

previous base written	next trit to encode		
	0	1	2
A	C	G	T
C	G	T	A
G	T	A	C
T	A	C	G

The result will look like this:

. . . TGCTAGCAGTCACTCATATACACGTCGCAG . . .

The encoding process is not finished yet. The string needs to be split up into overlapping strings with a length of 100 bases so that each string is overlapping another by 75 bases. To be capable to identificate the original file later (in case that more than one file is involved in the experiment), two identification trits are attached as well as 15 computed information indexing trits; both are translated to bases with the same method as the 100 data bases. The resulting DNA strand is now 117 bases long.



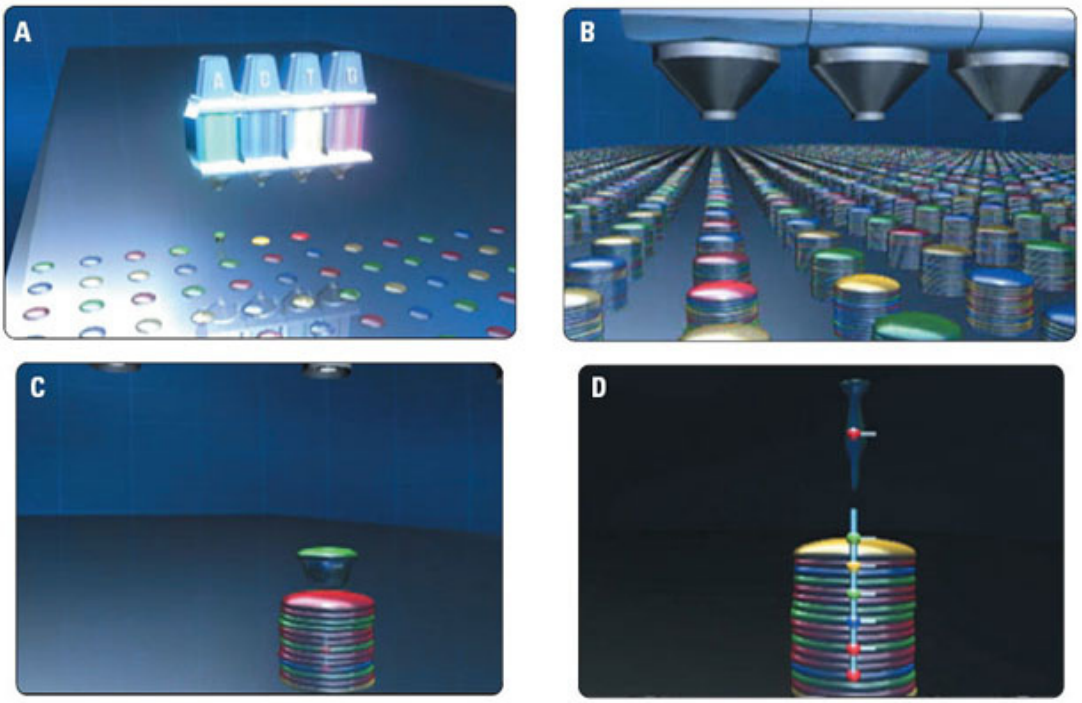
Schematic overview over the encoding process
Source: http://www.nature.com/nature/journal/v494/n7435/extref/nature11875-s2.pdf (20.4.2013)

To decode the data again, the encoding process is just reversed.

DNA synthesis

DNA synthesis has been an unfulfilled dream of genetic researchers for decades. Nowadays, companies like GenScript or Biomatik offer synthetic DNA on a very large scale. There are many different ways to synthesize DNA.

Goldman's team used an updated version of the so called oligo library synthesis (OLS) process of Agilent Technologies on the SurePrint microarray platform. In an anhydrous chamber, small amounts of phosphoramidites are applied with a Inkjet printing technique on a planar surface. This process is repeated until the oligos reached their desired length. Once the synthesis process is finished, the oligos can be detached from the surface.



A) The first layer of nucleotides is deposited on the surface
B) The oligos are growing after further layers of nucleotides
C) A new base is added to the chain
D) Close-up of the chain
Source: http://www.genomics.agilent.com/GenericB.aspx?PageType=Custom&SubPageType=Custom&PageID=2011 (20.4.2013)

Nick Goldman's Research

NY Times Interview

We tried to contact Mr. Goldman personally, but as we guessed, he is a very busy man and sadly could not find time to answer our questions. We therefore looked for an alternative and found this interesting interview which answered most of our questions. You can find the original mail we sent him in the Appendix.

Nick Goldman
Source: http://nygenome.org/sites/default/files/Goldman%20-%20450_0.jpg (20.4.2013)

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.

Our interview

As already said, we hoped to interview Nick Goldman personally. We sent him an email with our questions, but sadly didn't get an answer. Whilst browsing the Internet, we found a New York Times interview from the 28 of January 2013 in which Mr. Goldman answers some questions similar to ours. On that base, we tried to figure out the answers to our own questions.

The interview is based on the article published in "nature", like our own questions. The Interview gives you a good view on what the problem with today's data storage systems are and what Goldmans inspiration to change something was. Our questions were as follows:

How and when did you get the idea for this research?

Goldman had the Idea while sitting in a pub with his colleague Ewan Birney. They were talking about the problem they have in the European Bioinformatics Institute, because the institute is providing the world with some of the major biological databases: genome sequence databases, protein structure databases and others. This information is saved on huge expensive hard disk servers, but actually the information comes out of DNA sequence or protein. So they had an Idea. Why not use the DNA itself to save the informations? .

What were the major difficulties during the experiments?

The New York Times interview doesn't say much about what the difficulties in the experiments themselves were. But as we know from the article in "Nature", cost and a high error rate are the main problems.

When do you think will this method of data storage be used on a large scale?
Goldman says the 0.75 MB are already a relatively large scale when compared to other experiments on this idea. However he also says they have only just proved that it is actually working and hope to proceed with larger scaled experiments soon.

For the time being information storage in synthesized DNA is possible, but it's slower than other data saving methods. Do you think that this process will one day be easy, fast and cheap that it will become our new hard disks?
In terms of speed, the method will probably never be competitive with silicon based data storage methods, but then again that's not what they hope to achieve anyway.

You have proven that the principal of information storage in DNA is possible. What's your next step?
A very important step is to make the process cheaper. They hope to achieve this by using less error correcting codes and by doubling the information in the same amount of DNA. They also want to miniaturize the process to make it cheaper and more useful to the general public. They plan to achieve this by the use of microfluids.

Do you think it might be a problem for future generations that so much information is stored? A personal profile of all people could in principle be stored consisting, for example, of all actions such as internet activities, credit card transactions , webcam films?
This method is planed to be used as storage for high-value scientific informations. Personal profiles aren't included in those.

Discussion

The technique of storing and accessing data in synthesised DNA is proven to be possible. This doesn't sound very impressive, it but actually is a huge success for research on this topic. For the future Nick Goldman's team of highly qualified scientists has planned to achieve many different goals: The first and most important out of all, is to find a way to store more information in less DNA, mainly because the whole process of storing information is very expensive. By storing the same amount of information in half the DNA, the price could be halved. The team hopes to achieve this by using less error correcting codes.
We think biological data storage will be mainly used to backup already existing information. Important information such as scientific knowledge or historical informations and sources could be stored in bunkers together with crops which preserve the most important seeds for further generations (such a "crop-bank" already exists in Norway). Over a longer period of time, the storing of information in synthesized DNA might even become an alternative to hard disks used in our computers today. This is a bit of a long shot though, as todays technique is still very far away from something like that and it's highly unlikely that this bit of science-fiction will ever come true. But then again, if you remember that the very first computers nearly took a house to fit in and didn't even have the abilities of today's cheap "made in china" calculators, the situation might look very different in half a decade! Of course genetic engineering will always have objectors who think it is ethical problematic. However, we can't see any ethical problems in this technique whatsoever, mainly because the DNA used to store the information is synthetic and does not involve living creatures of any kind at any point. This type of DNA information is not "readable" by biological systems. Therefore we don't see many problems

(beside the technical ones...) in the storing of information in DNA and we are convinced that if this method would become easier, quicker and a lot cheaper, it would be one of the best suited ways to preserve rarely accessed data in archives all over the world. However, we do see one small problem, not with the technique itself, but with the users: If DNA based digital storage ever will become available to many institutes and governments, people might go completely crazy with the almost unlimited storage space available:
More and more of what we do is being monitored, recorded and saved somewhere. Therefore we are slightly concerned that too much unnecessary information could be stored making it impossible to handle and exactly as confusing as the data-chaos we might be drowning in in the nearest future.

Summary

In face of the rapidly growing amount of data, a storage in synthetic DNA could be auspicious for medium-term (ca. 50 years) data storage with only rare access needed, for example government records or to mention a scientific example the fastly growing collection of records of CERN's Large Hadron Collider which grows by 15 PB (15*1015 bytes) per year. However, the high cost compared to most other ways of archiving as well as the high error rates are problems yet to be solved. With the further development of DNA synthesis, the technique is not only supposed to become more economic, but also to last longer. It is certainly worth to keep an eye on the further evolution of this brilliant piece of genetic engineering.

References

New York Times interview:
http://www.nytimes.com/2013/01/29/science/using-dna-to-store-digital-information.html?_r=0

"Nature" 494 page 77–80 (07 February 2013)
Supplementary information to the "Nature" article
http://www.nature.com/nature/journal/v494/n7435/extref/nature11875-s1.pdf
http://www.nature.com/nature/journal/v494/n7435/extref/nature11875-s2.pdf
Huffman code:
http://www.ebi.ac.uk/goldman-srv/DNA-storage/orig_files/View_huff3.cd.new
http://www.minnesotalawreview.org/wp-content/uploads/2011/08/Zimmerman_Final.pdf

http://www.genscript.com/gene_synthesis.html?src=google&gclid=CJbE0b2L17YCFYuR3godywMAQQ
http://www.srf.ch/player/tv/einstein/video/schweizer-dinkel-fuer-die-ewigkeit?id=a291631d-eb58-4bcf-9130-3f2933eedd40
http://biomatik.com/Company/AboutBiomatik.aspx
http://www.genomics.agilent.com/GenericB.aspx?PageType=Custom&SubPageType=Custom&PageID=2011
Short animation about DNA synthesis:
http://download.genomics.agilent.com/platforms/microarray/SurePrint_Animation.mpg

Appendix

Our original e-mail to Mr. Goldman:

Dear Mr Goldman

We are a group of students from Gymnasium Kirschgarten Basel, Switzerland doing a term paper for our biology course. We are doing it about your research published in "Nature" in January on "information storage in synthesized DNA". For our work it would be very enriching if you or someone from your research team could answer our following questions.

1 How and when did you get the idea for this research?
2 What were the major difficulties during the experiments?
3 When do you think will this method of data storage be used on a large scale?
4 For the time being information storage in synthesized DNA is possible, but it's slower than other data saving methods. Do you think that this process will one day be easy, fast and cheap that it will become our new hard disks?
5 You have proven that the principal of information storage in DNA is possible. What's your next step?
6 Do you think it might be a problem for future generations that so much information is stored? A personal profile of all people could in principle be stored consisting, for example, of all actions such as internet activities, credit card transactions , webcam films?

Thank you for your time! We wish you much success in your further research and we are looking forward to hear more about this study.

Best regards
Pascal Grumbacher