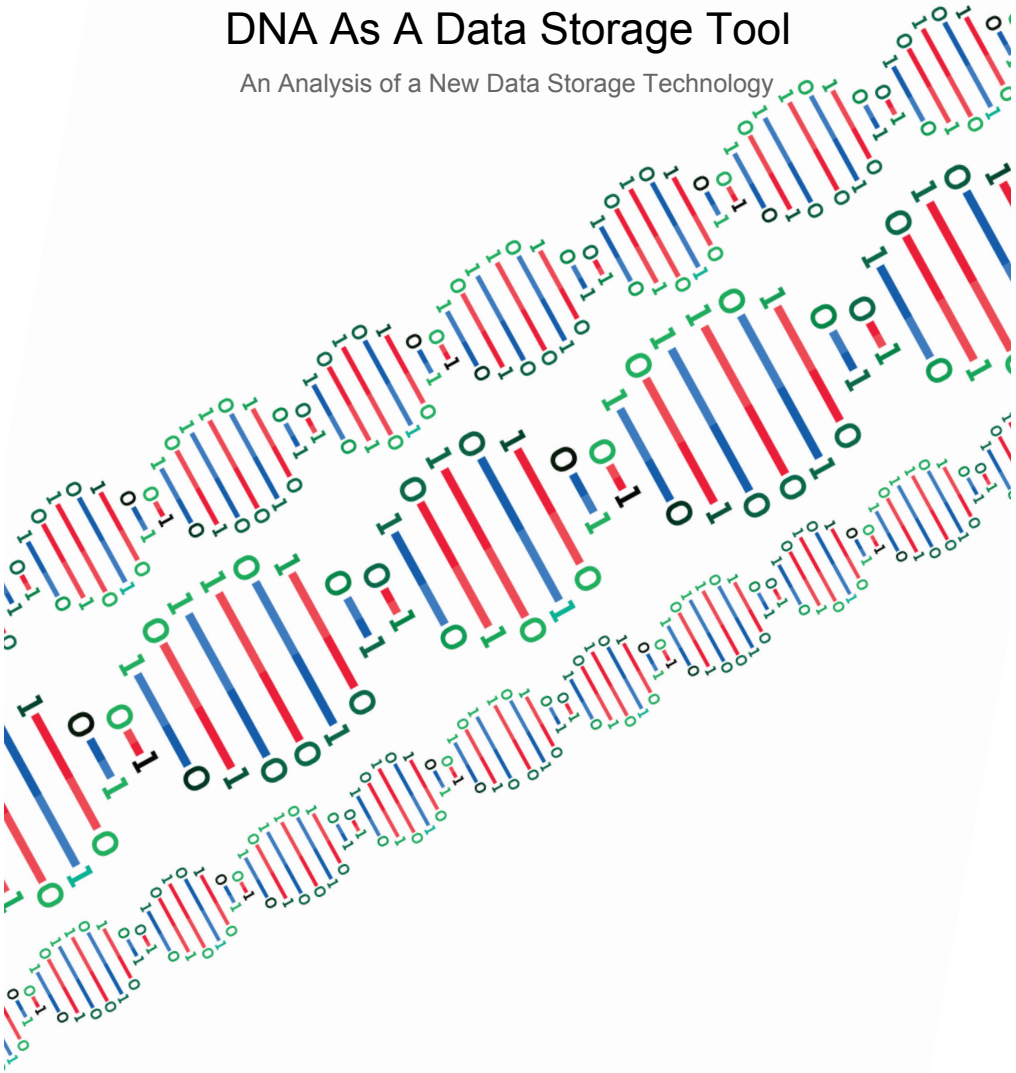
DNA As A Data Storage Tool

An Analysis of a New Data Storage Technology



Philipp Maier & Cristian Leonte
5B Gymnasium Kirschgarten
9/04/2018

Index

1. Foreword

The reason we chose this particular topic is because we are interested in seeing, whether, in this world where technological innovation is getting exponentially faster and computers are reaching their physical limits, DNA data storage will be the next big breakthrough in computing. As the rate at which we create new data increases, the world needs bigger and bigger data centers each year, and DNA data storage could be a perfect candidate to meet the demand of densely packed long term archive space. The possible switch from inorganic computing to organic computing that could come with this breakthrough is also something that interests us greatly.

Another reason for us to choose this topic, is our fascination with DNA manipulation, so we thought this paper would be a great opportunity to explore and find out more about what these tiny molecules can do. Paper, tape, vinyl, and hard-disks are all storage means that become obsolete after only a few years, but DNA can last way longer without losing information and it packs that information much more densely. Why not make use of the great tools nature has been using to store entire blueprints for organisms as small as *e. coli*, all the way to creatures the size of blue whales. And with an already existing need to decode and learn more about DNA, be it in medicine, archeology, paleontology or any other field, using the new things we learn and finding a new application for them seems only natural.

In this paper we set out to find out how DNA data storage really works, an just how commercially viable it could potentially one day become. To be able to conclude if it is able to be commercially viable, we will also look at the pros and cons of this new technology, the history and the future challenges it faces.

2. Introduction

Life has used DNA to store all of the information about the many countless organisms on this earth for millions and millions of years, and now we too are starting to take note of the potential storing power of genes.

It was in 1988, when researchers in Harvard translated digital data into a DNA sequence and back for the first time, and stored a small image of a Germanic rune on a strand of DNA.[1] The next big advances in the field however did not take place until 2011, when a team of bioinformaticians had an enlightening idea while pondering over the expenses and limitations of conventional computing methods.[2] Why not use the DNA, which can store the information for constructing entire organisms inside of cells, the tiny building blocks of life, invisible to the human eye, to store massive amounts of data in minimal spaces. The team, working at the EBI (European Bioinformatics Institute), knew that writing and reading such DNA data storages would take much longer than the conventional tools we use today. But it is not speed or ease of access which makes DNA data storage so attractive but instead the compactness and longevity this new method could provide.

Soon enough however, the team at the EBI, managed to set a new record by storing 739 kilobytes of data on a DNA strand.[3] With that interest and research, DNA based data storage will become more commonplace around the world, and with predictions that by 2040 the amount of data would far outgrow available silicon based storage capacity it would offer a great alternative for long-term archiving and large data sets for research of any kind.

[1] Nature, “How DNA could store all of the world’s data”.
[2] Ibid.
[3] Ibid.

3. Description of technique

The way we store data currently is by writing information to metal hard-disks as 1s and 0s and then translating that into legible data using binary, a base-2 numeral system.[4] A simple number that is only one digit long is stored as a combination of four bits (ex. 0001, 0010, 0011) and for every new digit that is added to a number, four new bits are added.[5] This only gets more complicated when letters or symbols have to be encoded.

The way DNA data storage works is by using the four available nucleotides (**A**denine, **C**ytosine, **G**uanine and **T**hymine) as bits in a string of data.[6] To store information in a chain of DNA, you first have to convert it to binary, then to ternary (a base-3 counting system) and finally translate that to the four possible nucleotides using a diagram like the one provided below.[7] While going through this process, the 8 bit long string was shortened to only 4 bits, which is 50% shorter. Please note that this is only one way of encoding information to DNA.

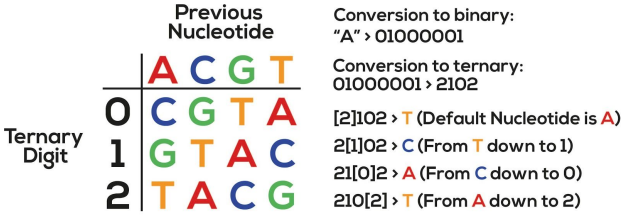


Fig. 1 A diagram used to translate ternary to nucleotide chains and vice-versa

To actually synthesize the nucleotide chain, a very complicated, long, and costly procedure has to be done. To write a nucleotide chain from scratch, one has to add new nucleotides to an open chain in a stepwise fashion and also in microliter volumes in order to avoid as many mistakes as possible. This procedure is still being improved on as it is still far from perfect and not commercially viable yet.

To read the new synthetic DNA strands one first has to multiply it many times. This can be done by injecting the encoded DNA into a bacterium, which will multiply the DNA everytime it reproduces. Before the strands can be read, some preparations have to take place. They first have to be cut up in pieces of increasing length so that each piece is only one base longer than the previous. After that, a coloring agent has to be applied to each strand according to the last base in the chain. Then the strands are inserted in tiny tubes filled with a conductive gel. Electrical current is then pumped into the gel, which will make the strands travel through the tubes to the other end. Since smaller pieces travel faster through the gel than long pieces, the strands arrive in order of length. The machine can then scan for the coloring agent that has marked the last base of each piece and note down the order in which the colored bases passed through.

[4] Explain that stuff, "Hard drives".
[5] Ibid.
[6] Leo Bear-McGuinness, "Is DNA the future of data storage?".
[7] Ibid.

4. Documentation of research institutions visited

As part of this paper, we visited a research institution and interviewed some of the people working there. For this task we asked several different laboratories across Switzerland and were granted an appointment, which included an extensive tour of the facility and a chance to take pictures and ask questions. This possibility was generously given to us by the Genetic Diversity Center (GDC) of the Department of Environmental Systems Science at the ETH Zurich and of course by the kind Dr. Aria Minder, the technical Director of the facility, who took time out of her day to give us a complete and comprehensive tour of the laboratory, and Dr. Niklaus Zemp, a bioinformatician at the GDC, who we interviewed.

During our visit we learned of three different ways of extracting DNA. The first technique works by mixing the sample with magnetic beads that specifically attract only DNA strands. The beads are then fished out, which pull the strands along with them. The second technique is pretty straight forward, the sample goes through a paper filter which has small holes through which most molecules will pass. DNA on the other hand is usually too big to pass through the holes so it stays on the surface of the filter paper. The last technique employs the use of chloroform to break the DNA free from cells and then uses the fact that differently dense mixtures will separate from each other, which makes it easy to separate DNA from the other contents that may be in the sample.

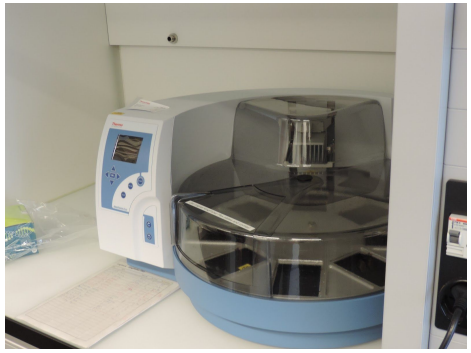


Fig. 2 A machine used in the magnetic beads technique to "fish" out the beads

Dr. Minder also explained to us that before any sequencing can be done, the purity and properties of the sample has to be checked. The purity can be checked using a device called Qubit fluorometer, which measures the amount of DNA, RNA or protein in the sample by looking at how much UV light is absorbed. To see how long the strands in a sample are, another useful machine can be used (Fig. 3). This machine employs the use of Agarose gel to group strands with similar lengths together. Based on where the strands end up it can output a graph that gives the expert handling the sample a good overview over how long the

DNA strands are. If the expert is not content with the length of the strands in his sample, he can use a tool that utilises ultrasounds to break the strands into shorter pieces.

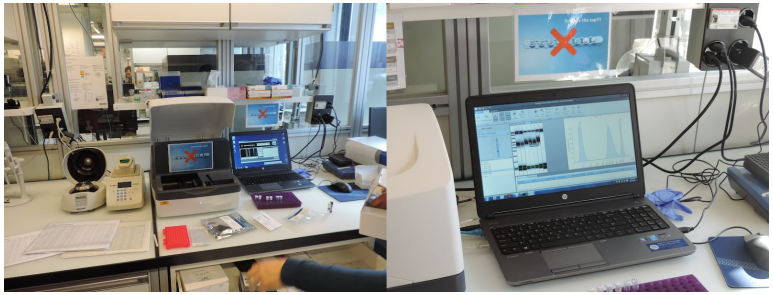


Fig. 3 Left: Machine used for Agarose gel electrophoresis ; Right: Generated data

The Genetic Diversity Center uses traditional sequencing, but also Next Generation Sequencing (NGS). NGS is an umbrella term for many new sequencing technologies. The GDC uses a Illumina MiSeq machine (Fig. 4) to be specific. While traditional sequencing employs the method described in “3. Description of technique”, the Illumina MiSeq employs a new and modern technique that takes 60 hours and can do 20 million sequences. It sequences DNA by first multiplying the inserted fragments of DNA creating many millions of clusters using a technique called bridge amplification.[8] After having multiplied the fragments, the next step is called ‘sequencing-by-synthesis’ where one nucleotide is added to the strand at a time, complementary to the nucleotide on the strand.[9] Each type of nucleotide (T,A,G,C) is fluorescently tagged, i.e. has its own color code. After one nucleotide has been added to each of the millions of fragments, they are all excited using a lightsource.[10] This makes the specific color code of the nucleotides visible. By analyzing the light given off by each fragment and saving the information, the machine thus reads the DNA fragments and catalogues them into data sets.[11]

[8] Illumina Inc., “Illumina Sequencing by Synthesis”.
[9] Ibid.
[10] Ibid.
[11] Ibid.

Fig. 4 Left: The Illumina MiSeq sequencing machine; Right: Generated Data

Dr. Minder also showed us a small device, no bigger than a TV remote, that uses another new technology called Nanopore sequencing. This device can read extremely long DNA strands but has a too high error rate to be reliably used in the lab. The way Nanopore works is by utilising a protein pore through which strands are pushing and read in real time. (Fig. 5)[12] The reading functions by measuring disruptions in the electrical current inside the pore. Because of this unconventional method, the Nanopore can read very long strands in one go and also requires less preparation.[13] A big downside to the Nanopore is that the device only reads the DNA and cannot interpret the data.

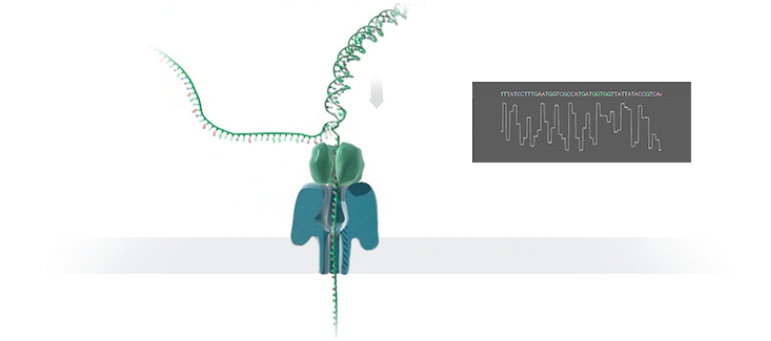


Fig. 5 Nanopore sequencing using the protein pore

After the very comprehensive tour, we got to talk to a bioinformatician, Dr. Niklaus Zemp. His job at the GDC is to interpret, analyse and correct data sets that are generated by the sequencing machines. He also gives tips and aid to the people that rent and use the machines for their experiments. We asked Dr. Zemp if new machines are getting better at avoiding mistakes and he told us that the machines *are* getting better, but not by much. He

[12] Oxford Nanopore Technology, “DNA: nanopore sequencing”.
[13] Ibid.

explained that to avoid mistakes one either has to train algorithms to find mistakes and fix them or you compensate the high error rate by sequencing more samples.

We also asked Dr. Zemp what big hurdles have to be overcomed to really improve the speed with which DNA is sequenced and interpreted. His answer was that it is computer processors that have to become faster so that they can analyse data sets faster. He also mentioned that many outdated systems and software are being used, which were made by biologists and not by big companies.

When asked if he could tell if DNA data storage will ever be a reliable technology, Dr. Zemp said that it is not possible to say just yet. He also told us that one always has to decide between slow but safe methods and fast but not so safe methods.

5. Discussion

DNA would be an incredibly space efficient and perfect medium for long term data storage. It has a information density which is 10^3 to 10^6 times larger than that of conventional tools, and can potentially store said information for centuries or, if properly kept, maybe even millenia.[14]

Today, many archives rely on magnetic tape to store seldom accessed files, since they pack data more densely than silicon does. DNA could be the next big step for such data storing facilities. It is faster to access than the magnetic tape, stores magnitudes more data per cubic-centimeter, and where magnetic tape holds information for little less than a decade, DNA could potentially remain readable for centuries. Numerically speaking, DNA could potentially store 215 petabytes (215 million gigabytes) in a single gram.[15] This would allow us to store the entire data on the internet today inside a shoebox.

However, before DNA data storage can become a viable alternative to the tools available today, there are many challenges researchers need to overcome and research in the field continuing, many problems have cropped up and are in need of solutions. DNA translation is prone to mistakes and harder to correct than conventional silicon data storage tools.[16] On top of that, accessing specific sets of data without having to decipher the entire strand proves to be an issue as well. The way scientists tackle these issues is as follows: The errors that occur when writing the DNA are not easy to actually correct, instead scientists make large amounts of copies of the same sequence, which make it possible to drown the random errors which occur while sequencing.[17] Another method used is one, where all parts of the data in a strand set always overlap with up to four other sets in the next DNA strands, which in turn makes it easier to locate and correct errors. To combat the problem of accessing specific files, researchers have used PCR (see Fig. 5) to multiply specific DNA sequences so that those could be extracted and read specifically.[18]

[14] Nature, “How DNA could store all of the world’s data”. In the following: Nature.
[15] Science, “DNA could store all of the world’s data in one room”.
[16] Nature.
[17] GenomeWeb, “Challenges of DNA Data Storage”
[18] Ibid.

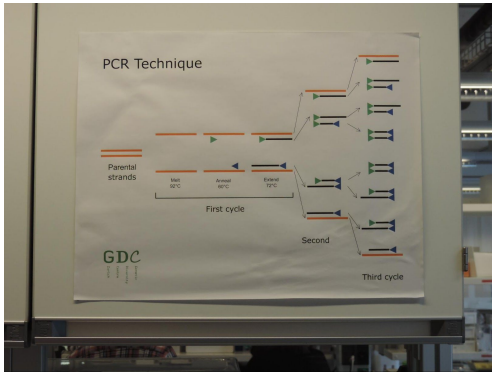


Fig. 5 poster explaining the polymerase chain reaction (PCR)

To the above mentioned difficulty of correcting the DNA strands and accessing specific data quickly, come the large cost and large amount time and money it takes to synthesize and duplicate DNA in the first place. The costs of generating and then reading a strand of DNA containing around two megabytes of information are around 7000$ and 2000$ respectively.[19] Add to this that current technology could only reach around 1.8 petabytes of storage instead of the aforementioned 215 and one can see why DNA data storage will not yet replace the silicon storage units we use today.[20]

[19] Science, “DNA could store all of the world’s data in one room”.
[20] Ibid.

6. Summary

DNA, which nature has used to store all the information needed for life on earth, has become the focus of much research over the past decades, and now some people think it could be the next big step in data storage. It can hold massive amounts of data in the tiniest of spaces and remains readable for centuries or more if properly cared for. Yet, there is no way that DNA data storage can be commercially viable at the current costs and time needed for production, rates at which mistakes appear and maximum storage space of just over 100 megabytes. However, with costs of DNA synthesizing having experienced a two-millionfold reduction since 2003 and the maximum storage capacity having increased by 250 times just between 2011 and 2016, taken together with the increased interest and with research teams at the EBI, Microsoft Research at the University of Washington and all over the world, many scientists are confident that DNA storage is a real possibility in the future.[21]

Nick Goldman, a group leader at the EBI went as far as to say that these improvements in technology are “very credible”.[22]

Not all scientists agree that DNA storage will become commercially viable anytime soon, maybe ever, but Goldman reassures us: “While past performance is no guarantee, there are new reading technologies coming onstream every year or two. Six orders of magnitude is no big deal in genomics. You just wait a bit.”[23]

However it may be, you will not carry your favorite song encoded in your cells and silicon valley will not change its name to ‘deoxyribonucleic acid’ valley any time soon.

[21] Nature.
[22] Ibid.
[23] Ibid.

7. Bibliography

Websites and articles:

- Digital journal, “DNA as a data storage medium: Progress and challanges” <http://www.digitaljournal.com/tech-and-science/science/dna-as-a-data-storage-medium-progress-and-challenges/article/497730> [21/02/2018]
- Explain that stuff, “Hard drives” <http://www.explainthatstuff.com/harddrive.html> [04/02/2018]
- ExtremeTech, “How DNA data storage works” <https://www.extremetech.com/extreme/231343-how-dna-data-storage-works-as-scientists-create-the-first-dna-ram> [21/02/2018]
- Genome News Network, “How does DNA sequencing work?” <http://www.genomenewsnetwork.org/resources/whats_a_genome/Chp2_2.shtml> [04/02/2018]
- GenomeWeb, “Challenges of DNA Data Storage” <https://www.genomeweb.com/scan/challenges-dna-data-storage#.Wsp8JtNuaoh> [21/02/2018]
- Nature, “How DNA could store all of the world’s data” <https://www.nature.com/news/how-dna-could-store-all-the-world-s-data-1.20496> [21/02/2018]
- Science, “DNA could store all of the world’s data in one room” <http://www.sciencemag.org/news/2017/03/dna-could-store-all-worlds-data-one-room> [21/02/2018]
- Science, “DNA Fountain enables a robust and efficient storage architecture” <http://science.sciencemag.org/content/355/6328/950> [21/02/2018]
- The National Center for Biotechnology Information, “DNA Synthesis, Assembly and Applications in Synthetic Biology” <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3424320/> [04/02/2018]
- Wikipedia, “Artificial gene synthesis” <https://en.wikipedia.org/wiki/Artificial_gene_synthesis> [04/02/2018]
- Wikipedia, “DNA Digital Data Storage” <https://en.wikipedia.org/wiki/DNA_digital_data_storage> [04/02/2018]
- Wikipedia, “DNA” <https://en.wikipedia.org/wiki/DNA> [04/02/2018]

Videos:

- Illumina Inc., “Illumina Sequencing by Synthesis” <https://www.youtube.com/watch?annotation_id=annotation_228575861&feature=iv&src_vid=womKfikWlxM&v=fCd6B5HRaZ8> [01/04/2018]
- Leo Bear-McGuinness, “Is DNA the future of data storage?” 2017 <https://www.youtube.com/watch?v=r8qWc9X4f6k> [04/02/2018]
- Oxford Nanopore Technology, “DNA: nanopore sequencing” <https://nanoporetech.com/applications/dna-nanopore-sequencing> [5/04/2018]