DNA DIGITAL DATA STORAGE

Leoluca Bernardi

Felix Miolin

Luca Cannavó

Gymnasium Kirschgarten

Table of Content

1. Preface

When we found the topic of DNA digital data storage our interest was immediately piqued as we all have an interest in modern and developing technologies. The dream of immense storage capacity and incredible longevity seemed too good to be true, captivating us throughout the project.

We recognized the enormous potential of this branch of biotechnology and were curious to know more. To us, it was especially interesting how DNA digital data storage presents a potentially perfect solution to a problem that has been ever more present in technology. Furthermore we were curious to discover first baby steps of a new field of scientific interest.

Reading into the topic we had many questions, the main ones being how long it would take for this Technology to develop into something applicable in everyday use and what was keeping us from already having researched further into this field.

2. Introduction into DDS

2.1. Context

Ever since the first computer was built scientists have worked on optimizing the storage of data. Speed and capacity were essential in the 80's, since then digital data storage technology has come far, from a Megabyte being the size of rooms to having several terabytes on a chip the size of a fingertip. This development has also been followed by a rapid growth in the generation of data in need to be stored for prolonged periods of time. In 2012 the total information storage of the entire world was around 2.7 ZB. This number is said to increase by 50% every year, which would mean that the total information storage of the world today would be 70 ZB, assuming linear growth, which is most likely not. Many facilities, digital services and people depend on the storage of large amounts of data for future use, which raises a problem.

All current mediums of data storage have objectively short lives and are very susceptible to corruption or damage. Modern memory cards and chips are sustainable for approximately 5 years, classic hard drives have an even shorter estimation than that in addition to being very prone to damage from moisture, scratching or heat.

Even though current solid-state drives, SSD's, operate more safely and better, they are also at risk for damage if not power-driven for several months. This has led researchers to investigate new mediums of data storage, that are more resistant and sustainable under the conditions. Eventually DNA was taken into consideration as a storage medium for digital data.

2.2. Area of Application

Since DNA is already able to store the entire blueprint of the human body, the idea is that through manipulation of the strands it is possible to write upon a DNA strand like you write data upon a scratch disk in a hard drive. DNA is extremely durable and can survive through the harshest of environments In addition to this, DNA has been found to have an enormous storage potential next to being extremely small. DNA has a half-life of around 500 years, which means that natural loss of data would only happen after 500 years, at which point only half of the information will have degraded. It is estimated that a single gram of DNA can hold 1 Petabyte of data, that is 1’000’000 Gigabytes of storage.

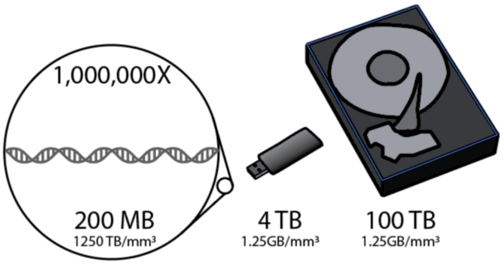


Figure 1: Storage capacity of different mediums

Due to these reasons, DNA is by far one of the greatest candidates for future data storage devices. There are other candidates for this, though objectively none of them have a potential matching that of DNA digital data storage. With further research this data storage form could even be modified in a way that it can be incorporated into a living organism, allowing the stored data to be renewed and essentially creating a self-sustaining storage device.

3. How DDS is achieved

The process of storing and reading digital data on a string of DNA can be divided into two general parts. The process of writing and the process of reading. Currently, it is much easier to read DNA as almost every laboratory has the means to sequence DNA with relative ease. Writing or editing the DNA to hold the wished information is the greater hurdle in the development of DNA digital data storage. For the process of storing data on DNA two things are necessary: A means to synthesize DNA with a wished sequence of bases, and a Coding language that allows for binary code to be translated to bases and vice versa.

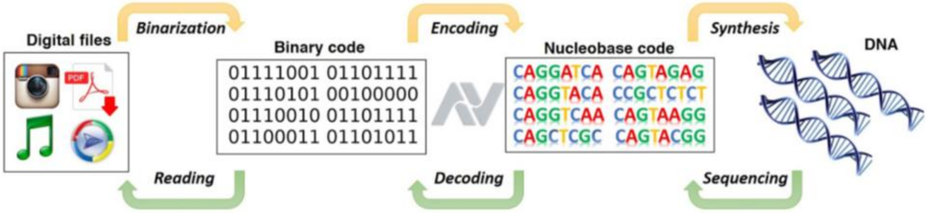


Figure 2: Flowchart of storage process

Currently, the synthesis of DNA is very expensive and has a relatively high error rate, making it necessary to reduce the length of strands wished to be printed, as a single mistake will render the entire strand unreadable.

There are many different coding languages allowing for the encoding of digital data on DNA. None has yet to become an official norm and scientists around the world still use different methods to encode different kinds of data onto DNA.
One of the simplest encoding methods called: “Simple transcoding” is very simple to understand but not very efficient in terms of Base count-efficiency.
Another method is the Huffman code. This method is much more base-efficient and used in actual transcription processes. Both these methods and two others are visualised in *Figure 3***.**

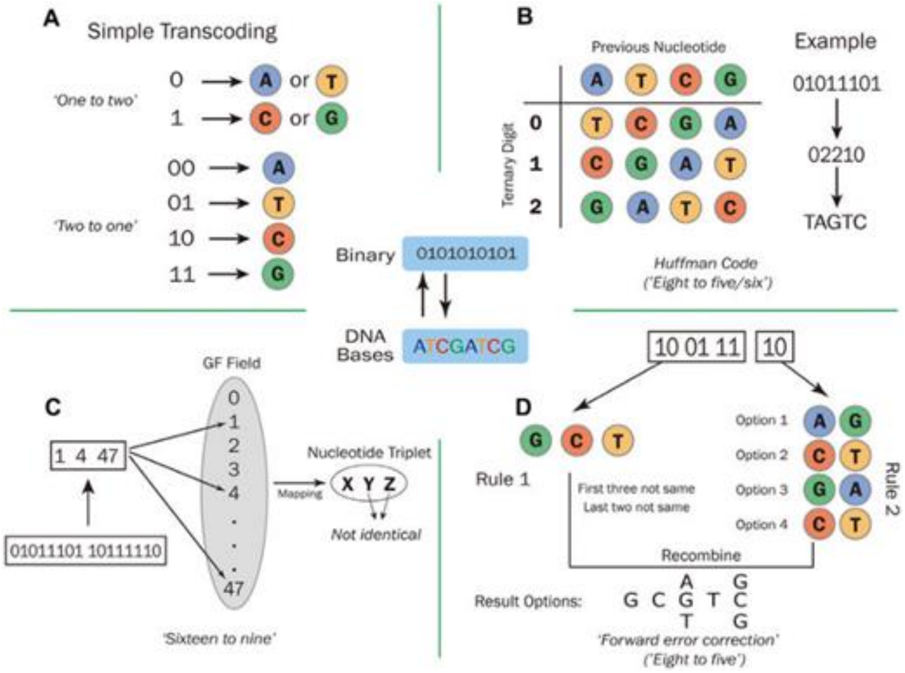
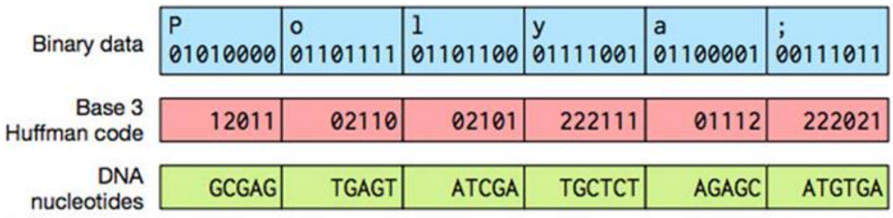


Figure 3: *Encoding methods*

For example would the encoding of a short word in binary, with the help of the huffman encoding structure look like this:



Binary data	P 01010000	o 01101111	l 01101100	y 01111001	a 01100001	; 00111011
Base 3 Huffman code	12011	02110	02101	222111	01112	222021
DNA nucleotides	GCGAG	TGAGT	ATCGA	TGCTCT	AGAGC	ATGTGA

Figure 4: Example of the Huffman code

With the first, simple transcoding method it was very easy to encode data onto DNA. As mentioned earlier, this led to the written DNA sequence being very long, which caused problems during the synthesis of the DNA strands. Because of this, scientists kept the length of the DNA sequences to around 150 base pairs. This was not too short to hold too little information but neither too long to have a sub-par succes rate during synthesis. Scientists developed a method of being able to write lengthy bits of code split onto many of these 150 base pair long sequences of DNA. While encoding information onto these DNA sequences scientists use the first few Bases to write a marker that assigns each string of DNA to a certain spot in the entire sequence.

This makes it possible to order the DNA sequences during sequencing by arranging them with the help of these markers, creating a long readable sequence of DNA. The longer the full sequence is, the more of the 150 base long sequences are needed and the longer markers are needed to be able to effectively sort the strings. With the help of logarithmic addressing, the length of these markers can be consistently kept to a minimum.

4. Swiss Institute of bioinformatics

We reached out to the SIB to ask our questions regarding DNA digital data storage, as we had seen, that they were currently conducting research and have in the past too, been very proficient in the Field of DDS. A group leader by the name of Christopher Dessimoz at the SIB, who was also a professor at the University of Lausanne held an interview with us.

3.1. Interview

How would you describe DNA-DDS with your knowledge and background?

“The storing of information is a central problem in computer technology. Humanity produces a lot of data, we need a better storage type. Still today tape is a long-time storage medium. This is a very stable method compared to many other methods like flash drives or disks. DNA is attractive for long-time storage because it is extremely small, extremely stable and it is digital. It is like a binary system with four possibilities, A/C/T/G. If we store data on DNA, we won’t have any problem in the future reading it. Because as long as humanity exists, we have DNA and we just can read it with a DNA sequencer.”

How much does the process cost and how will the price develop?

“DNA sequencing is getting cheaper every year. However, DNA synthesizing is not getting cheaper as fast as sequencing. If the costs are high, people are using something else. But at one point it will be competitive, it will make sense to produce and a new market will form. It is a matter of time and it is not going to be more expensive than it is. Too long sequences of DNA can cause an error for the whole batch, but if you know the error, you can deal with it. We can deal with different errors if we know them. There is already a market for DDS today. There are a few companies that can store data on DNA for people. In a paper in 2013 was shown, that it was already competitive in long-time data storage.”

Where do you think it will be applied, which companies would use it?

“It is interesting for data centers. It is really hard to talk about the future, I can tell you about the present. I think there will be a lot of applications we will use DDS, but at the moment we don’t know. It is more probable that we use it for long-term storage, which we don’t need to have access very fast. Because it takes time to sequence it. If you think about pictures, some of them we want over two or three generations. If we store them on DNA, they won’t get lost and probably you never watch them, but you could always synthesize them.”

What is the direction this field is developing into and what is currently being researched?

"There is a new method to combine artificial DNA with your own. Then you can use your own as key to decrypt the data. But there is a high-security risk because you spread your DNA all the time. It is easier to get DNA from a person than we think. There are a lot of people that research new codes to store information, that are more stable."

What different encoding algorithms are there, and which is the most popular one?

"At the moment, the procedure is not used frequently and no standard method for the encoding process has developed yet. This occurs naturally as the field progresses."

How does the process of synthesizing Information into molecules work?

"If you imagine a disk image, the data is always encrypted in zeros and ones. A naive way to store data is, to use A/T for zero and C/G for one. Then you synthesize the molecule and at the end you sequence it and you got your zeros and ones back. But we can do better than that, 00 is an A, 01 is a T, 10 is a G and 11 is a C. With this method the string will be much shorter. Also, it is no problem that you need to store your data in small pieces, you just number your pieces. There are other ways to number your own DNA, you search identical parts that are overlapping, but this is likely to fail because you can have identical pieces at a separate place, then it is getting difficult. It is much better if you store the address. You don't need much space to store the address, this is favorable."

5. Discussion

DNA was recognized very early for its potential as a storage option for digital data. DNA is not only an attractive option for data due to its longevity and high durability but also because of its enormous storage capacity per space occupied. If DNA data digital storage is further optimized, it is set to replace many other long-term storage solutions, although there have been problems and limitations regarding the encryption and decryption of DNA. Many of which have been fixed, like the bottleneck of DNA synthesis not allowing for consistent, longer string to be fabricated and having a high cost. Scientists found a workaround, by marking strings with barcodes to specify what position in the code each string belongs to. This allowed them to attach many DNA-strands to each other, making the first problem negligible. Still, since no proper market for the synthesis of DNA has been able to form, the costs remain high. Though, as discussed in the interview, it is only a matter of time until this market will begin to develop, which will result in lower prices for DNA synthesis, like the current growth of the DNA sequencing market.

DNA also shows a lot of promise in the secure encryption of digital data, as DNA is extremely difficult to decrypt without knowing the given primer sequence of the encrypted DNA. As a proof of concept, it would even be possible to use your own DNA as a key for sensitive personal data. The DNA sequence of every human is unique, the problem is though, that it is permanent, so if one gains access to your DNA it is impossible to change your own DNA sequence to another to ensure further security. With arbitrary keys, this is different and is very attractive to consumers who seek to store large amounts of sensitive digital data on long-term storage devices. Data centers for companies like Facebook could be massively reduced in size and improved in safety if DNA digital data storage were used.

Another branch, scientists have been researching, is the incorporation of digital data into the DNA of a living organism, like bacteria or even a tree. This would potentially allow for the survival of the data through even harsher conditions due to the adapting nature of organisms. If it were possible to ensure the safe transfer of the data from generation to generation, the lifespan of DNA digital data storage could be infinite. Though preventing mutations in the generational transfer is a hurdle on which scientists are still working on. A large question on this topic is wether the procedure of inserting artificial DNA into a living organsism is ethically acceptable. The use of an organisms as a storage medium is a controversial thematic.

Currently, the largest problems for DDS are the high costs of DNA synthesis, the low retrieval speeds/rates, and the relatively slow process of encoding data into DNA, where no standard has been agreed upon yet. Many different encoding languages for binary into bases exist but each fulfills a niche purpose. Scientists around the world are using different methods of synthesisy the ones discussed in this paper are just some of many. All of these things only make the medium attractive to large companies wishing to secure sensitive or critical data, banks could use DDS for the long term and secure storage of personal details and as mentioned, large data centers could reduce their size or increase their effectiveness by uploading their data onto DNA.

6. Summary

In this paper, we explored the young field of DNA digital data storage and worked to understand the advantages and difficulties the researchers and scientists have faced. As the topic had quickly captured our interest we had not experienced much trouble finding a suitable field of study. This niche branch of biotechnology concerns itself with the benefits, the application and the process of using DNA as a digital storage medium. DNA is a highly attractive storage unit as it is space-efficient and durable. Additionally, it might be possible to integrate synthesized data DNA into a living organism effectively creating a self-renewing storage option. The hurdles come from the slow process of writing or synthesizing DNA.It is currently quite faulty and not as reliable as one might wish, but according to our correspondent this is not as significant of a hurdle for research in comparison to market accessibility.

We had contacted the swiss institute for biotechnology who arranged a meeting with an expert in the field, Christopher Dessimoz. He lent us an ear to our questions during a lengthy interview which we experienced to be a both engaging and informative way of learning about the topic and we feel most thankful for the opportunity he had given us. Afterward, we concluded that our expectations for the technique as a storage method might have been a little optimistic, but we were pleased to find that given enough time the concept might develop into a long-term storage option for state secrets or research projects.

7. References

- DNA Data Storage in Perl; Ui Jin Lee, Seulki Hwang, Kyoon Eon Kim & Moonil Kim; access date 19.10.2020; URL: https://link.springer.com/article/10.1007/s12257-020-0022-9

- Digital Data Storage Using DNA Nanostructures and Solid-State Nanopores; Kaikai Chen, Jinglin Kong, Jinbo Zhu, Niklas Ermann, Paul Predki, and Ulrich F. Keyser; access date 06.01.2021; URL: https://pubs.acs.org/doi/abs/10.1021/acs.nanolett.8b04715

- New Trends of Digital Data Storage in DNA; Lei Chen; access date 06.01.2021; URL: https://www.hindawi.com/journals/bmri/2016/8072463/

- Molecular digital data storage using DNA; Luis Ceze, Jeff Nivala & Karin Strauss; access date 06.01.2021; URL: https://www.nature.com/articles/s41576-019-0125-3

- DNA digital data storage; Wikipedia; access date 06.01.2021; URL: https://en.wikipedia.org/wiki/DNA_digital_data_storage

- Natural Data Storage; Manish Kumar Gupta; access date 14.01.2021; URL: https://www.researchgate.net/figure/PCR-based-alignment-based-and-primer-library-based-encoding-model-is-shown-here-Fig_fig4_277023595

- DNA and the Digital Data Storage; Lichun Sun, Jun He, Jing Luo and David H Coy; access date 14.01.2021; URL: https://www.hsj.gr/medicine/dna-and-the-digital-data-storage.php?aid=24516

- DNA-Encoding (ft. Christophe Dessimoz); ZettaBytes, EPFL; access date 15.01.2021; URL: https://www.youtube.com/watch?v=yPpCaiLiVQs&ab_channel=ZettaBytes%2CEPFL

- Future Computing: DNA Hard Drives | Nick Goldman; World Economic Forum; access date 15.01.2021; URL: https://www.youtube.com/watch?v=tBvd7OSDGgQ&ab_channel=WorldEconomicForum

- An improved Huffman coding method for archiving text, images, and music characters in DNA; Menachem Ailenberg, Ori D. Rotstein; access date 17.01.2021; URL: https://www.future-science.com/doi/full/10.2144/000113218

8. Image Sources

- Figure 1: access date 03.02.2021; URL: http://sitn.hms.harvard.edu/flash/2019/the-computer-science-behind-dna-sequencing/

- Figure 2: access date 03.02.2021; URL: https://ashutoshviramgama.com/dna-data-storage-synthetic-dna-future-of-storage/

- Figure 3: access date 05.02.2021; URL: https://www.researchgate.net/figure/Binary-transcoding-methods-used-in-DNA-based-data-storage-schemes-A-One-binary-bit-is_fig1_333908428

- Figure 4: access date 05.01.2021; URL: https://www.semanticscholar.org/paper/A-DNA-Based-Archival-Storage-System-Bornholt-Lopez/7b06ba3effa9fc7b2f194a355bcb69601ef1ea56/figure/4

9. Attachement

This is the complete Interview, transcribed by a speech-to-text program. It is not edited and bears no relevant information whatsoever.

Interview between Christopher Dessimoz, Felix Miolin and Leoluca Bernardi

Okay, excellent. So this just remind me the context your

this is not that's not like a bachelor worker. It's just like a term project or what do you call it? It's like a big assignment we have to do enter last year of biology. Okay. Yeah, exactly. Yes. It's a term paper. Again. So you decided to do it on DNA storage specifically, or is it part of a bigger theme? No, it's DNA data storage in specific, specifically, because we were like researching different. Well, we had to choose something and biotechnology and DNA digital data storage just really spoke to us.

Sure, so I just got distracted, because I'm running a Class A practic.

And I have it at university. There's a first year course with 160 students. So they've got assistance. In zoom, actually, now, on occasion, they have an issue, they asking me some stuff so that I get some distract. But this is something I can do. Okay, that's fine. Just that, you know, I'm It's not that I, I usually really need to be so distracted that I just need to have a quick look if it's something relevant. Okay, so

yeah, how can I help? Well, you probably saw the questions we sent you.

In essence, three of them can probably be

asked this one question in itself, but

would it be final? Who started off with the question

of how you with your background would describe the field of DNA digital data storage with your knowledge? and expertise? Yes. So

I mean, the, the problem of storing the information is, I mean, I think it's fair to say, it's, it's a really central problem in, in computer technology in general, and particularly the long term storage,you know, the humanities producing so much data, that, and instead of more or less an exponential growth, that we need constantly to find some better ways to store these data. And then there's, there's been lots of different attempts. And but you may, I don't know if you're aware of this, but still, today, the for long term storage, the preferred way is to do it, using tape,tapes, they degrade, and they need to be read and rewritten every couple of years.

And that's already viewed as being fairly stable by by compared to many other storage technology. Things like flash drive, or, or Well, I mean, you guys probably didn't grow up with, well, you still use disk. I mean, you use optical disk, I suppose, like in your PlayStation ormaybe in the car still. But you know, this already.It's also not lasting for that long, you know, they oxidize, and thenfairly quickly, you use use, I mean, certainly, at this scale of five to 10 years, you need to, to renew this. And then if you are going to things like SSD and hard disk, it's even less, less stable. So so long term storage really requires something and then, but the for that DNA is really attractive, because it's got quite a few properties that makes it a unique one. It's extremely small. I mean, in terms of like, the information that you have, you know, every cell in our body contains the entire in all the genetic data, as you know, as you've learned, a whole genome. So that's about 3 billion base, you know, and we have two copies from one from the Father, one from the mother. So that's like 6 billion base that is stored in every cell. And we have been on the order of 10 trillion cells. So that's really if you think about it, it's a huge amount of information is really, really quite small. I mean, I Can't remember, but I think I did some calculation once and you can probably find these online. But if you take the DNA that is contained in every cell, and you unpack it, well, if you just take it from one cell and young packets about two meters long, and then if you take for all of these cells in the button, you put it end to end, you can go all the way to from the Earth to the Sun, right? And then go back, and then you know, a few times, it's just come completely crazy that the length, so yeah, so I mean, it's just gives you a sense, like, it's really, really dense.

Another thing, which is very nice, which is really stable. Why because it contains, I mean, it has evolved to be really stable. Life could only you know, actually there is some there are some theories that we there's still a bit of uncertainty about what was before DNA. But some people think that maybe there was an RNA world. And so, but the RNA was not as stable, you know, it has a lot of nice properties also, as you know, to carry the information, the genetic information, but not as stable as DNA.

And again, you know, stability is, yeah, so you know, the proofreading, you know, you have a T the T pairing, the C and the G and so, you can enable as well, enables you to correct, sometimes a mistake to identify some mistakes, and then correct them. And so that's nice.

So it's table is extremely dense. And there's also one thing that is also quite attractive as a way to store information. And it is that well, okay, it's also digital, if you think about it DNS is, is digital, you know, it's not, it's, so that's a little bit also like computer technology, right? So, but instead of having a binary alphabet, zeros and ones, like in your computer, you have four possibilities, AC, T and G, okay, but it's still, you know, digital in the census, as opposed to analog, where you have continuous values. So that may, you know, that's also quite suitable that this this information is in this form. And, and then there's also an argument that, if we store information in DNA, then we won't have any problems to read it in the future. Because as long as we have intelligence forms of life, you know, we will, we will likely still have our DNA and, and as we know that they're all of the living being have the same type of DNA, so you can just use a normal sequencer, and relax. So you know, all of these things, they make it like a DNA, like quite an attractive medium. And then what has happened in quite recent years is that it's become really cheap to sequence. So then, you know, to read the whole, yeah, you know, so if you do a round of sequencing in biology, you may get hundreds of millions of reads, so the little bits of, you know, AC, T and G that are maybe a flank about 150 to 200 nucleotides in just one run. And I mean, I've done all the prices, I mean, usually, the machines are really expensive, but then you do lots of rounds on them, and then PR around, it may be cost, I don't know what it is maybe 500 francs, or 400 front, and you get 100 millions of reads, you know, each so you can enter the end of information, and then the costs go down and down every year. Actually, if you look over the past 15 years, the costs have gone down faster than Moore's law. I don't know if you know, Moore's law. This isn't even about that. I heard of it. But I've never really thought Moore's law. It's a law that was formulated by I think, maybe the founder of Intel, you know, the the CPU and that they, he noticed that with the technological improvement, they could double the number of transistor that they put on a chip, every, every 18 months, so you can double and, and so while that's really quick, and it's this is what created the computer revolution, and actually that's Still, some people say well, more, they sometimes then equated this to like the speed of the computer doubles every 18 months, which is, to some extent, the two things that correlated, that's not been true. But the density keeps on, you know, like, you look at the latest, the latest like Apple chips, and, you know, they use they every transistor is is, you know, a few, I don't know, maybe 30% smaller than the previous generation. So, you know, if you every time you reduce by 30%, it's like an exponential process, upgrade. So it still holds many, many years later. And we DNA, the decrease has been even faster in terms of sequencing. So that has really changed things. So the first the cost of sequencing for the first for the human genome around the year 2000. Well, it was maybe from 1995 to 2000, roughly speaking, they estimate that the cost of sequencing was more than $1 billion. And you see, now it's just a few 100. Franklin's can probably they're probably some some people who can sequence a human genome for $100, if you read too many of them, you know, I mean, don't quote me exactly on this value, but you can probably check, find online a source, and you will cost and you can show how it has decreased. But if already at the cost, I was watching your videos. And the big point was also the cost of synthesizing DNA, right, which is one of the biggest obstacles you said, right? Because in 2017, you said it was like about 10,000 francs for a mil four mega gigabyte megabyte. Yeah. I don't know what is the latest figure. But you're right, because this is purely market driven. If you have enough people using a certain technology, then there will be more competition, and then the price can go down. And for for synthesis, the price have not gone down as quickly as we're sequencing. There are some people who are working on this and I saw just a few months ago, some colleagues from ETH Zurich actually they published a paper on, on on using that technology that you have in crystal liquids display LC you know, like a projector. And the projector, the way it works is that it's, it's, yeah, let me tell you just before the way it works, the synthesis for the leading technology is a bit like an inkjet printer, you have like a matrix, and then it just it just the poses like either AC T or G. And then there's a reaction that that gets incorporated. And so that's how they grow. But that technology, what they do is to selectively choose which base to add, they use a crystal liquid, you know, display, I believe. And so in the crystal liquid display, you have for each pixel, you have a way to control how much light goes through, right. So if you want to have something black, you block the light. And that's why you know, this display not perfect, some of them, you can still see you know, it's not black, it's like dark gray, because there's some light that goes through. But you can do that selectively for each pixel. So now imagine you have now again, your matrix, and then by blocking, sometimes the reaction you will only going to and let's say you have a chemical reaction that is induced by lights, then you are going to be able to say okay, on just, you know, we put some free a soft, free floating aid, and it's only in this position that we want the A to react and then to be like gated in it now, does that mean the consistency of the printing would be safer, more stable? Because I also in the video, you said that the problem is that you can't have too long sequences of DNA, because of errors that would cause the entire batch jail. Yes, that is also that's also the case with this technology. But you know, this, and I may I may have mentioned this also in the video, that's not a problem. It's just an engineering problem. I mean, you're you're you have a certain error rates. And as long as you know what this error rate is, and you can control these error rates, you just deal with it, you know, you just have some some redundancies and error proofing mechanism. As long as it's not an unexpected, you know, level of errors. There's no problem. In fact, your hard disk, you know, you look at your phone, don't think that the storage doesn't have any error. It's got lots of errors, but it's been done. So that you know there are some control some checksums and when the when the controller of the storage node Are we've got a bad apple here, they stop using it, etc, you know, there's all these layers on top of that, but it's, it's full of mistakes. It's just that it's, these are mistakes that I control. So if you need to generate a perfect string of DNA, a perfect molecule of DNA, that may be quite tricky a few phrases you need for genetic engineering. And then you may not like to have a technology that has a high error rate. And by the way, also, just a little aside, if your technology is really cheap, but it has a high error rate, you can also just produce more, and then get rid of all the ones that contain errors, and then keep the one that Okay, all right, yeah. So that's kind of a bit of a brute force way of decreasing your error. So the error, we can work with this with the error is just, but you want to also see what is the price, okay. But just to go back to your, to your point, it is true that whatever technology you're taking, the costs are not going down as quickly as we thought they might go down. But I would say, you know, these are the type of things, it's really, you can have a tipping point, you know, as long as the costs are high, people are using something else. And all of a sudden, you know, that becomes competitive. And now it makes sense for companies to start producing this, you know, the market grows, etc. And then you have a replacement of the technology that happens very quickly. So this type of technology replacement, they, it's, it's not a linear progression, you kind of have slow adoption, and then all of a sudden, you have like, an inflection point, and then, you know, big, big, big growth. So we just need patience, essentially. Yeah, I think so I think it's a matter of time, it's just a matter of time is not going to get more expensive, and then it's going to get more attractive. And, and then the costs are going to go down really, really quickly. Because in terms of the technology, you know, there's nothing that you need, you don't need to go and mined uranium or, or very rare made metal or or you don't need something that is the the size of a factory to produce DNA, you know, this can be done with relatively small parts. And so the big costs are also in the in the in the research and development, but then when you get the scale, you know, the cost per unit is going to be really cheap. This is why for instance, in this plug here, which is just the power adapter, these are a little computer in there, you know, there is a chip, you know, to build this chip to build this chip, maybe you have to spend a million for a billion francs for the for the factory, but then you produce so many of them, that per unit is quite cheap. So you have this, this phenomenon, with with this technology. But one thing I want to say though, is even though it's not mainstream, there already is a market for DNA storage. Today, there is already some things that are done. With dentistry, some companies first something they come in, they say, okay, could you store this information in DNA? And so, of course, these are like pioneers, you know, it's a very special thing. So what could it be? Well, it could be like a publicity stunt, you know, these, like Massive Attack, which is a band, I don't know, they've had like the 20th anniversary of their, when the album came out. And they say, we're going to store this in DNA and like makeup, make a bit of a splash, you know, and they approach our colleagues. And, and, you know, they paid what it costs, and they did it. What we showed in the nature paper already back in 2013 was that it was already competitive in scenarios where you want to store the information for a very long time. Right. And it's really for a very long time. You know, it's going to be cheaper than just reading and writing the tapes. Many times, you know, if the DNA is intact, you just have to wait. So I think there are relatively few people who are worried about storing information for 2000 years. Yeah, right. But what kind of people are concerned about that, in that case, they may be, for instance, governments for some information that are really important, the location of nuclear waste. I mean, there's some information that is really valuable that you want to store in the long term. But again, you know, I'm just telling you, for this type of very specific application. There's already a market today it's very small. But then when the price go down, there is a point where, you know, if you think about the amount of information that is out there, you know, every bank But also store archives about all the transaction, you know, legal application where you need also to, you know, I mean, I was involved in a lawsuits, you know, where I was an expert witness, the amount of paper that is printed by by lawyers is just is just incredible. You know, they they have mountains of documents and so yeah, for these type of things, I mean, I think you will, you will see that, you know, at some point DNA is going to be like, the most competitive Okay, yeah. Because also like, Facebook thing, like, add up bytes or exabytes of storage? Oh, I'm not sure. But, but but but I'm sure it's very large. Can you just wait for a second? Let me just have a quick look. Yeah. If there's any problem. Okay. Okay, good. Sorry. It's just It's nothing, nothing urgent. Yeah. So yeah, I mean, they want one other thing. For instance, you know, the, you know, the NSA in the US. This is the agency, which has all the information spies, you know, that's all Yeah. Remember, maybe Edward Snowden and revelations there. So they obviously want to store a lot of information. They want to store your emails, no. Yeah. And they have, they have facilities, they have places data storage center.

And I can't quite remember, there's some public information on this. That's the scale is also staggering. Here. It's like a fight somewhere. And it cost them $1 billion, you know, and they can they can store a certain amount of information. They're also really interested in DNA storage. Yeah. So yeah, right, this would come to a question like, Where do you think it would be applied? Like, what companies will in the future also be especially interested? Yeah, what areas of life? Well, I mean, data centers of friendliness is also quite interesting base. You know, for instance, in Amazon, if you're using Amazon, not as a consumer, but as a, as a company, you know, many companies use the Amazon data centers, there are a certain type of data that you store there, this is for long term archival, okay, it's just a different type of data. And the cost there is very low, as long as you don't access the data. And then if you want to access the data, it's more expensive than when it's on disk, presumably, because they put it in a in some media, there are, there may be another as efficient to retrieve the data, but which are cheaper in the long term. So that you know, they are they right there, they already have a product today, that would be maybe quite suitable for, for for genius, or it's called Amazon glacier. If you want to look at glacier, click Show. Yeah. And so you think primarily, it will be used for data centers?

I don't know. I mean, look, I don't know about the future. No, I mean, it's very hard to tell about the future, I can tell you about the presence, whether it is used, I can tell you about some trends. I mean, I think there will be lots of applications that we cannot foresee at the moment. It's clear is that what is going to be compelling sooner, or for long term storage where you don't need to have access so quickly, because it does take some a bit of time to sequence. Uh huh. Yeah. So for instance, you know, we are having this conversation, these videos going live, and you've got, you know, there are some you need some memory locally, like buffered information, you know, all the pixels, they don't come at one so you have to reconstruct the image and then store this on some temporary and that, I don't know if it will ever be DNA, I don't think doesn't make any sense. But, but but for instance, wedding picture, think about wedding pictures or, or, or pictures from your, from your childhoods, you know, think about your parents. Where do they store this information? Physically, generally? Yeah, but why did they do that physically? While they don't need to access it that quickly? Right? And also, they probably don't trust that if it was just on the computer, what if the computer crash and the backup and the visible? Or were they okay? Yeah. What if, what if there's the you get a hacker who encrypts your disk and then ask your, you know, some money to access it, or you do, you do a mistake, or maybe you know, so this is a real product. And this is something that you want to store more for them for a few years, you want to get it ideally, for your lifetime, perhaps even the lifetime of your own children and grandchildren. At some point, maybe, you know, you don't, people also don't want to necessarily look at all the pictures they had, you know, even from the next generation, but over a lifetime, you know, it took about easily 50 100 years, 150 years web, if you then also think about the lifetime of just you know, over two to three generations. And then it's not so crazy, to spend a bit more money to have it in a place, which maybe you never access. But at least you can access if you want, you know, it's you put it in a in a safe in the back, and you know, it's in DNA, that's, that's still gonna be readable. It's like your backup plan. You don't have to do anything, you don't have to reread and restore, making sure that, you know, you download the latest driver that the new version of Mac OS is still going to be able to open that file. This is a real problem. Yeah. It's terrible. But you talk about like, you're not so sure about the future. But you know, what's, what's in the press? And like, what's possible? So, like, what do you What's the direction DNA data storage is going in? Like, right now? Are we are you researching how to more efficiently encrypt, like, translate data onto a piece of string of DNA? Or where is the field right now? I mean, I can tell you about the work we did. And I was not the main author on this, I helped a little bit, but it's really main credit is, is my my colleague, Robert grass, who's really quite a star in this field. And Roberts, what he did, and we had discussed about how do we create information? Yeah, because, you know, you're thinking about Switzerland or so Bank Secrecy? And, you know, how could we do that? That's an interesting subject. There are some, some some data centers, you know, for sensitive data you got away from, from St. akia, kind of an idea that was a bit more like a stunt. And he was thinking, Well, you know, we can combine their, you know, we can encrypt information using standard technology. But then the key could be your own DNA, the natural DNA. So that's quite cool now, because you combine kind of artificial and natural DNA. So essentially, you have the sample, and no one can read it. But then you come, you spit in it, which is mixing it also with your own DNA. And then now we can sequence both get the information and then use the information from your own DNA to encrypt to decrypt this. So it's kind of a fun thing that actually in this paper, we describe that you can read the paper, I can send you the reference, if you like, Oh, yeah, please. We encrypted Turing's the Turing paper thing that describe how to crack the Enigma machine, you know, from the Nazis. World War, which was a kind of classified for a long time. And so that's stored in DNA. And then the key that we use is we use some part of DNA that have variable length. I don't know if you've learned this in biology, but this is what is used in for paternity tests.

For forensics, you know, in crime scenes, yeah. primers that amplify some region that are known to be very and it gives like a barcode like a signature. Yeah, we actually just did that like two months ago. Okay, well, so you can we can use some regions that are variable that you know, unlikely to be the same for you and for Leo is your name or Leo Luca? It's Leo, Luca, Luca, okay. So it may be you know, different And then if you have enough, entropy, this is the measure of randomness is the measure of information, if there is, if it's sufficiently distinct, you know, if there are just two variants, you know, like, it's like, either you have brown eyes, blue eyes, well, you know, you can try one or the other, and then you'd be able to decrypt it for there's not enough enough randomness inside. But if we take some enough characters that are not correlated, there are distinct, there are many combination. And we can use this to, to, you know, as as the way to as a key to decrypt the information. And then the paper describes it. So but honestly, this it is a bit of a stunt, because it's kind of a cool story. It's a cool piece of work. It's thought provoking. And okay, we did it. But actually, you probably would not want to have the key in your DNA. Why would you not want that? Well, I would want it but you wouldn't use it that much. Probably. But what's what's the security risk? Yeah. Well, I don't know. You can use a loose DNA DNA quite easily. Right?

Yeah, you could go and then try to, to get a comp and then get a hair, you know, from your toilets. You're shedding DNA all the time. Right. So your key is like, and also, of course, you know, there's the perfect twins, they have the same, same key, you know, so you don't want your your evil twin to go and get all the money from your, your secret bank account. Right? So that's also not so great. I mean, the thing is, there is a I mean, you know, that with the going and getting your DNA secretly, that's a real risk, you know, there was a story, I don't know, if you, you followed that, but nowadays, more and more information that is stored in things like 23andme, did you know about these type of products? Yeah. And their sample, and then they get genotypes. So this is not the length polymorphism that gets that gets read it's their variable region in the genome that are known and, and people are just checking, you know, with chips, okay, do you have a C or T under position, you have an A or gene, that person that gets recorded. And then you can look at, you know, maybe, you know, your ancestry and some other, you know, the risk that you have of prostate cancer and our case, it's giving some in for some trades, we can, it's quite predictive, others is a bit harder. But there are big databases of that. And there's one guy he got caught, he, you know, he murdered some I can't remember exactly the story, but I think, you know, for 20 years, or 25 years, he was not found. And then what they did is the found a crime, you know, so of course, they have from the crime scene, they have the DNA. So they knew what they were looking for. And then they find someone in the database, which is predicted to be like a cousin. Okay, of the, you know, some, there's some further income. And then what did they do? They went and kind of cross checked with all of the people who live in that region, you know, who could possibly be a cousin and so on. And then they had the suspect, you know, somewhere, or I don't know, maybe more than one, but they had someone they thought, okay, you know, that looks like, you know, that could be your man, you know, it's like an ex policeman or I don't I don't know how exactly what was the problem? And so but then how can you prove that? What they did is they follow this person, and they got the cigarette butts. You know, what's left of it? And they sequence DNA from that? Yeah. Fascinating. Maybe you should change your topic that's, like, crime scene research with DNA. That's, you know, the point is that so, you know, people if they want to target you and get your DNA, it may, it may be easier than you. Yeah. And then also, the thing is, for instance, if you get your your credit card store, what do you do? You block the blockade, and you get a new one. How do you how do you get in?

Yeah. Once it's out, its out. Right. I mean, there are also some other issues.

But yeah, I mean, that stung. I mean, it's just like a display of what it can do a potential Right. Yeah. So yeah, that's right. That's a that's an that's a showcase. And they've been people who, who develop a lot of research around finding more efficient codes to store the information because If you if you if you are, you have a better grasp on the on the type of errors that arise in DNA, then you can be more efficient and more robust in how to store the information. A lot of cool ideas, you know, they're using very advanced technology for information storage. That's actually a big question we have right now. Like, you need to translate the binary into the DNA with a CT and G. Yeah. And right now said, there are different codes for that, like, what different ways are there, right? Because we don't really understand when we try to do research, because there's so much different, so many different ways of doing it, and like mortal so one of the main, like, is very, like, most popular method right now. I mean, most of you know, the everything mark, every application is can I mean, it's at the moment is so pioneering, you know, it's it's only much later that you will see the standards emerging. So you cannot say what is the most popular approach, but it is, you have some general, I call them general purpose, there are some things that have been developed for other purposes that have been adapted for DNA storage, and they work very well. And yes, you can, you can do some research to find some way to improve by 10%. Maybe by 20%, you can optimize, you know, how much DNA Do you need to store, you know, so that you don't, you know, it's not wasteful, because, you know, if you can get away with 20 molecules, why would you synthesize 100 molecules? You know, it's just more, okay. Actually, I should spread that I mean, well, no, that's not that's true, it's more expensive to do to do more, even though, of course, DNA can be duplicated very easily, right, you know, that you can use PCR DNA. Yeah, it's a template. So it's very cheap, actually, to, to copy. But you know, it will also cause them some errors. And then at some point, still, it's good to know how much DNA you need, really, for the, so you can, you can optimize the protocols in all sorts of way. And this is the nature of research that people, you know, it's not like in a company where you really have one goal, and everything is going to be towards, like, that bottleneck, you know, which is like the main cost center, for instance, in research, we are interested in all these different aspects and who knows what will be the bottleneck, but maybe one of these things can turn out to be really important. This is the nature of basic research is that you don't know exactly where this is going to be to lead you in terms of the application. But you do this because you've identified some problem that you can improve upon. I only have another five minutes or so. So please, if there's Is that okay, or you had lots of other questions, or? Well, it's fine. No. Okay. Well, then I would have like, one last question, because the process of yeah, oh, yeah. It's cathartic. Okay, the process of synthesis to information storage, how, roughly how does it work? So you've synthesized the DNA, and then you somehow had to? Well, you use a code? Yeah, two very basic code, a very basic code is that if you have 0000, you look at what you have on your hard disk, okay. Or let's say, a disk image, okay? When you take a disk image, and you just look at those zero, and the one that it may, it's made of you, this is understandable to you, you can imagine, like you, you will know, you can even use like a tool for that that is going to give you the information that you have, if you're using Linux, maybe it's the tool is DD or one of these tools that is really low level, and it's telling you what is stored in each of these bits.

Okay, you take that. And so this is a string of zero and one. And you understand that this could encode anything, right? a PDF, and mp3, a movie, a text, any file that you have on your computer, you don't need to have a different code for the different type of file. At the end of the day, all that has been already sorted out in your computer, and you just need the zeros and ones that you know, describe all of this data. That's clear to you. Yeah. And so if you gave me the zeros and ones and I just I was I had this really amazing memory, I could be the storage right and then like in 20 years, come back and say crystal, what is the what is the information I start and I tell you 0110, I give you the string, back euro, you write these back in your computer. And then you know, lo and behold, you mounted that did your disk image, and you've got all of your files. So the job is just to be able to store the zeros and ones and then to give them back. How do you do that in DNA? Well, if you have four possibilities, A, G, a very simple way to do that is to say, I'm going to store I mean, a very naive way is to say, I'm going to use a or T for zero, or C and G four, one. And now for any zero, you, you, you, you store it either as a or is it T, and for everyone C or G and so then you store it, you have your DNA, and then I mean, you synthesize this molecule. And then you read it back. Okay, now, can I simplify a little bit because we already said that we cannot have like a super long molecule, but let's just let's just assume that you are really rich, and that you you can order a molecule of any length and if someone is going to produce it for you, you just have to pay enough money. Get very expensive if it's very long, but now we just paying out using money to sort our problems. Okay, your story is long molecule of zero and a CTG and and then at the back then at the end, you sequence it, and all the A, you turn them into a zero and all the all the empty with zero, C and G one, and now you've done. Now, we can do a bit better than that, what you notice is that we're actually since you're forced days, maybe you can use them to encode actually two characters, right? You can do 00 is going to be a 01 is going to be a T one zero isn't cheap. And one one's a C, for instance, if you do that, now, if you have a string of length 1000 you only need a piece of DNA of length 500. And then how now how do you deal with the fact that you have to you can only write in small pieces? Well, that's okay. You know, you just number your piece. A little bit of the piece is used to number the piece. It's like a barcode I think right? marker. Yeah, you can call it how you want to know in which position you have. There's another way which is what you do when you have real DNA which is you find some overlap. You know, you have thing that overlap, but that will fail. Sometimes Sometimes, you know, if you, if you have, you can have ambiguity, you have two pieces that are identical, and they're in two different places. And then it starts getting really difficult. But it does overlap with this region or with this region. So it's much better to store the address. But you know, what the address? The beauty is that and that you will learn if you go to university, I don't think you know, I don't know if nasm is the type of thing they teach you, but the amount of information that you need to store the address is only grows logarithmically. Yeah, right. Out of information, you only add a linear amount. That's very favorable. So this is why for instance, you know, you think about how many people can have a phone number with, with, with one digits? How many people can have one? No, one digit, you get 10%? Right. About my heart about two digits? Well, 99 plus zero, you know, if someone can take 00, right, yeah. 100? And then how about three digits? Well, it always gets more than it's 400. That 1000s? Right, three digit 1000s 10 1000s 100,003. See, every time you multiply by 10. Okay, you just add one more position, you multiply by 10. So that part's the address, it costs you a little bit of space. That is not too bad. Yeah. I think he said it's something like 20 colons long, generally, you need to ask how much information you want to store in bands, you can optimize this, maybe that's what we use, I can't remember. But well about the coding technology, then, like the coding language, if you can call it that, if you have any recommendation where we should look into what's like, the modern like, I mean, there are loads. I know, I mean, if you want to describe this in your paper, these Reed Solomon, that's a really a well known standards. And you can probably find something on Wikipedia. So use also to store information in CDs, and three, wonderful Reed Solomon, there is fountain codes, which are also quite fancy, which are used for broadcasting. But I think you better just explain one well, in your in your opinion. And actually, I can tell you how Reed Solomon works. I mean, some of these things, the way they work is silent. You do some some, you know what is a polynomial in mathematics? Yeah. Yeah. So you do things like a x squared plus bx plus c? Yeah. Okay, so this is a polynomial. And so if you're looking for this equals zero, you know, how many solution they are, if you have just a polynomial of degree one, so it's just a line? How many solutions do you have? Usually? If you have a function, How many times will it cross? 01? Only once? Right? Yeah, if, if it's a quadratic, so two terms, twice, and then twice? Usually? Yeah, I mean, you can have something that we have no solution, but usually you have to, okay. And so, and if you have, if you have three, then you have, you know, three terms, then you have you have 03, and three pieces. And so what happens is that if you have usually three, observation, then you can fit a polynomial of the attribute you have, I just don't want to make you don't say anything. Intuitively, if you're talking about a linear relationship, you know, it's a x plus b, you only have to find your a and your B you to piece of information, right? And then you can do your system of equation and you get a and b. Okay? If you want to store three pieces of information, you can store it in a polynomial of degree two. And so now you have a, b and c as your your parameter. Now, the trick is, how do you make it more robust and then redundant to save some information is you use a polynomial of degree, let's say three, which has four values, but then you provide six values. Okay? It's overdetermined. So that means we do any of this, you know, four out of the six values, you can reconstruct the parameter of your polynomial. If you're missing one or the other one, it doesn't matter, because he's over determined. And you've got that there is no constraint, you see, this is a little bit away. So when you store this extra information, and it doesn't matter which one you're missing, you can always reconstruct the information with a subset. And then you can even do it a bit more clever. So where even if there are some mistake, not only missing data, but mistakes, then you can still correct for it, etc. So these are quite some clever thing. But it's almost another paper, you know, it's almost something about the coding theory, which is fascinating. But it's, you know, maybe maybe you can just give us a glimpse, but guys, I really go Okay. Okay. I think you've got a lot of material. Yeah, thank you very much. Sorry for keeping you this long. No problem. My pleasure. And well, it was really nice.

Good luck, let me know, let me know what happens. And if you could send that paper because I would be very interested in reading it immediately. Okay, thank you. If I forget. Wonderful. Okay, bye bye.