

# DNA-Storage and Quantum Computing Application

## A very interesting and completely new scientific direction

K. Bertels<sup>1</sup>

<sup>1</sup>University of Ghent

June 25, 2024

### Abstract

This paper proposes a new approach to build a scalable way of doing in-depth analysis of the large amounts of digital data we produce on daily basis. It is a combination of existing and still growing digital data, quantum computing applications and a DNA-based way of storing information. The content of the paper is to sketch the state of the art in Quantum Computing and will introduce the huge amounts of existing and future digital data. The paper also presents also a new full-stack for any new DNA-data storage device. This full stack is based on the published DNA-data storage Alliance but makes the explicit link with quantum data, digital and DNA-data storage. Such a DNA-device already exists but this paper makes the link between digital data, quantum results and DNA data storage. There are still open issues such as operating temperature and operational speed. Important to know is that temperatures around -5 Celcius are needed and no extremely low temperatures such as millikelvin are needed to store DNA-based information. Evidently, DNA-data storage is also applicable for pharmaceutical and medical purposes.

## 1 Introduction

The computer hardware technology is reaching the limits of the size of CMOS-transistors, used to make computer chips. Some authors call this the information catastrophe. [1] The current size of such transistors is at 2nm, and once they become smaller than 1nm, they enter the quantum mechanical universe, with direct implications. This universe makes them behave in completely different ways such that quantum elementary computations are done in a probabilistic rather than deterministic way. [2][3] A second important challenge is to store all the digital data we generate on a daily basis. Not many humans are aware of the fact that the CMOS-transistor size cannot provide enough storage space for any kind of classical and digital data. This is where DNA as a storage technology pops up. The goal of this paper is to introduce the complete innovative way of doing scientific research, by focusing on DNA-based storage technology to save huge amounts of digital data we have on our planet. We will also explain how this device can be combined to store computing or experimental results that can be further analysed using quantum computing technology. It is all very experimental but very needed.

Deoxyribonucleic acid, commonly known as DNA, is the fundamental molecule of life, encoding the genetic instructions used to develop and operate living organisms. DNA is composed of molecules called nucleotides, which include a nitrogenous base, a pentose sugar, and phosphate groups. These nucleotides are arranged in sequences of four bases: adenine (A), thymine (T), cytosine (C), and guanine (G), which are the core components that store genetic information

This paper first introduces the state of the art in quantum databases and talks about the link with DNA, by introducing the basic concepts of DNA, known as the most important medical data. DNA can also be used as a storage device. I then describe existing research about DNA storage and what the main approaches and advantages are. While I was working at Delft University of Technology, my last PhD student and now a friend-colleague, dr. ir. Sarkar, developed a fully operational quantum genome sequencing (QGS) algorithm. [4][5][6] <sup>1</sup> We had to deal with the main limitation, existing for any qubit technology, that it can only address small genetic datasets. Therefore, the number of qubits had to be limited, this is also the reason why we introduced the concept of Quantum Computing Logic or **QC-Logic** where we can abstract away from any error at the quantum physical level. Very interesting

---

<sup>1</sup>Interested in DNA Encryption, read [7].

is that we are now working on vaccine development, using mRNA. [8] The current paper presents how that QGS-algorithm could be connected to a quantum accelerator the results are presented. The paper concludes by presenting a possible path forward where we combine DNA storage with QC applications.

## 2 State of the Art in Quantum Computing and DNA-Storage

As stated in the introduction, our planet is going towards a new scientific revolution in the way that data are generated, stored and analysed with the new computing platforms we are trying to develop world-wide. Based on CMOS-transistors, there are initiatives towards 3D-CMOS chips, zero-energy computing etc, but none of these seem to provide a scalable and reliable way of intensive computing. In terms of building computer platforms, there are two main aspects one needs to look at. The first is the computing part, giving birth to many approaches and quantum computing (QC) is clearly a very high-potential approach. Important to understand is that the need for new quantum applications are needed to make the technology progress. [9] The second is the storage part. There are data-centers where huge amounts of data are being stored, and used for any kind of application purpose. It is well-known that, for instance, Google and Microsoft are planning on building new data-centers on either Iceland or somewhere in the ocean. The main problem these companies are solving is cooling the storage devices to a temperature that makes it less expensive for these companies to offer them. We already wrote quite extensively about the QC where the big problem is still that too many qubit technologies are in competition with each other, each trying to outperform the competition in building a stable and scalable computing platform. [2] Any adopted qubit-technology only generates a very small numbers of physical qubits that can be used for computing purposes. The suggestion we make in [2] is that it will take another 10-15 years before one or two of the QC-technologies will mature. As the focus of this paper is more on data storage, I will not explain a lot about DNA-based computing but will refer to some of their aspects later in this paper.

### Data Storage

In the same context of building a completely new computing platform offering computing and storage capabilities, a short overview of **quantum databases** is given. Figure 1 is based on the quantum computing papers produced so far, and the following papers give an overview of where we are in that field : [10][11][12][13][14][15][16][17] [18]. As QC is still in its very early phase, no usable results are available but maybe one of the first papers published on quantum databases is in 1997 [10]. Another

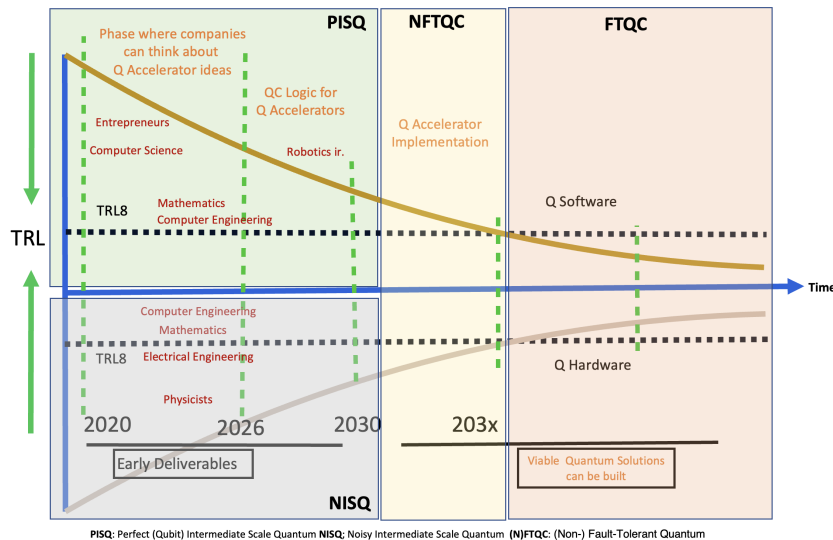


Figure 1: QC Vision figure

attempt was a paper published in 2009 and described in [11]. The authors provide a theoretical framework to extract information out of a quantum database, based on Bernstein and Vazirani's parity problem. In many cases, the authors make sometimes assumptions on where we stand in the QC-community. For instance, in [14] [15], all the authors refer to an exponential growth of the numbers of supported qubits but a good reference in that paper cannot be found. Paper [16] reduces the problem to such a small

size that it is not certain how that will scale to realistically sized problems. A similar approach can be found in [13], where a lot still needs to be developed to become operational. In addition, today, the production of high-quality qubits is an unsolved problem. Sometimes, the topics are so specific, as in [11], where the authors focus on how to construct search algorithms in quantum computer and duality quantum computers. Again, the relation with data stored in a database is not defined in any classical database term. Several other papers emphasise a lot the qubit states that are generated and used when executing a quantum algorithm. Features such as superposition and entanglement lead indeed to more amplitudes but storing those elements during the execution is not what we understand to be a database. Final comments are found in [17] [18]. In [17], the authors describe the practical implementation of theoretical principles for searching data in a quantum database. A similar comment can be made as the scalability of the approach is still a big question mark. A much more stringent comment can be found in [18] where the author explains that there is an arms race for building the first and best quantum computer, without even knowing when that technology will have matured enough to be used in practice. Not to turn any lecture of this paper in a very negative perception on quantum computing, the author

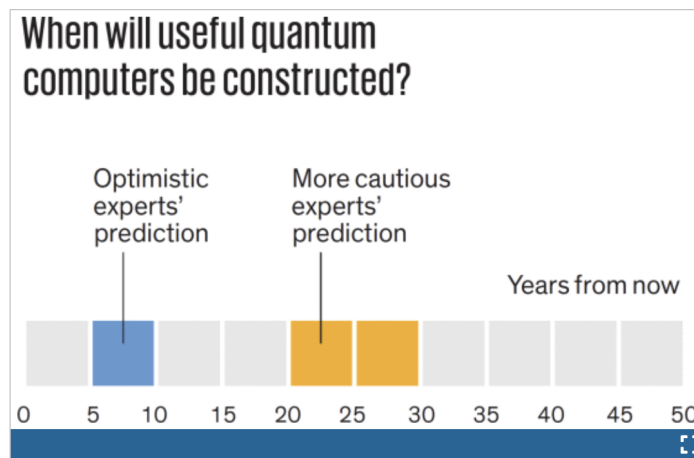


Figure 2: Quantum physics researcher from Yale

formulates his opinion as follows: **"Having spent decades conducting research in quantum and condensed-matter physics, I've developed my very pessimistic view. It's based on an understanding of the gargantuan technical challenges that would have to be overcome to ever make quantum computing work."** The paper where the quote is taken from was published in 2019 and Figure 2 expresses the most optimistic and the more cautious prediction. From 2024 up to 2030, we should see the first operational quantum machines on the market, and maybe it will take another 10 to 15 years. My personal experience on quantum computing is quite similar, but formulated from a computer engineering point of view and not that of a physicist. Hence the focus on DNA for the most part of this paper, but the link between quantum computing logic applications and DNA-storage is formulated. The exploration of DNA's potential for computing and data storage has been an area of active research, leveraging its ability to compactly store vast amounts of information. One of the earliest conceptualisations of DNA computing was proposed by the Russian physicist Neiman in 1964. He envisioned the radical miniaturisation of computing elements to molecular and atomic scales, foreseeing the use of molecular structures for data storage and computation. Significant advancements were later made by the American computer scientist Leonard Adleman who, in 1994, developed the first computational model using DNA, famously known as the TT (Test Tube) computer.[19]

While DNA computing is still in its developmental stages without significant achievements in computational power, it is increasingly recognised as an exceptionally potent medium for data storage.[20][21] In [22] and also mentioned by the American International Data Corporation, it is predicted that by 2025 the global data storage demand will grow to 175 ZB or  $1.75 \times 10^{14}$  GB by 2025. When comparing it to DNA-storage potential, one gram of DNA could store up to 215 Petabytes of data.

Figure 3 gives a complete overview of existing memory approaches, of which most of them are based on CMOS-transistors and the lowest layer represents a DNA-based memory layer. Figure 3 shows the different memory layers we will need, either for making a DNA-computer or a DNA-based storage platform. It is difficult to predict what computing and storage technology will make it to the market but it seems clear that quantum computing and DNA-approaches are either complementary or in competition with

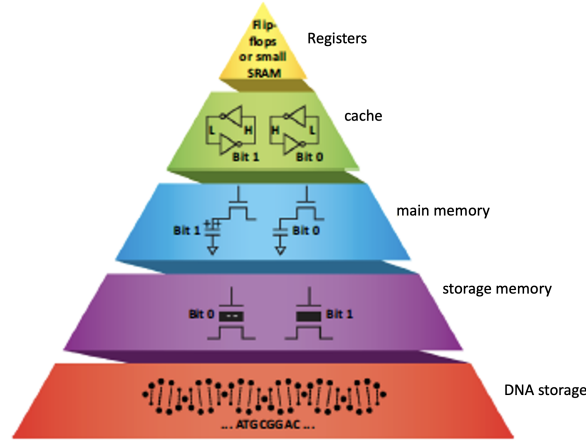


Figure 3: Full stack of memory components, including DNA

each other. In addition, we can be sure of one thing. Researchers world-wide working on DNA-related technologies have already shown that DNA can be used both for computing and storing data, in ways which are beyond comparison of classical approaches. There is no need to prove it is possible, the main challenge, however, remains to change that it can be done in a very fast way.

Humanity is producing so much data with all kinds of classical hardware devices, that we do not know where to store, leave alone compute, all those pieces of data. However, we can be sure of one thing: researchers around the globe have demonstrated that DNA can be utilized for both computing and data storage, surpassing the capabilities of traditional methods in remarkable ways. The feasibility of DNA-based technologies is well established; the predominant challenge now is to enhance the speed of these processes to meet practical, real-time needs.

### 3 Genome Sequencing and Analysis

Before diving in the DNA-storage aspects, we first describe the way the medical and pharmaceutical world is using DNA. We first briefly sketch the way currently the world deals with the use of genetic information that can be extracted out of DNA.<sup>2</sup> We use Figure 4 to explicitly present how the medical world uses, more and more, genetic analysis for medical diagnosis and medicine development.

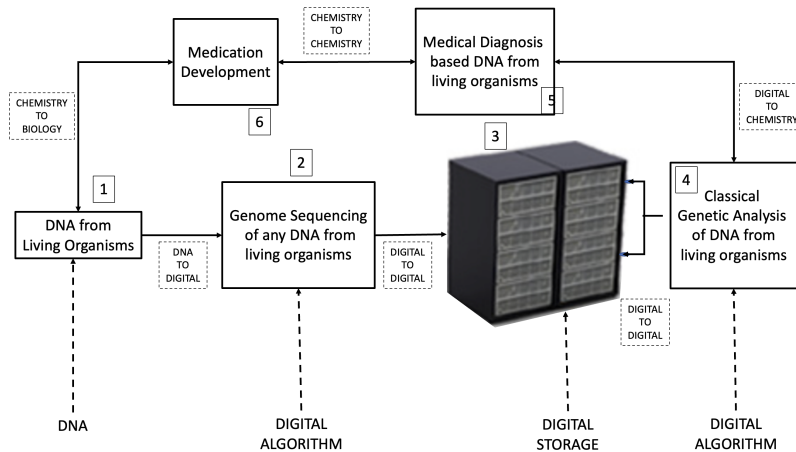


Figure 4: DNA and Genome Sequencing

1. **DNA of any Living Organism** - Every living organism on this planet is sharing the same basic DNA. Certain properties create the enormous diversity we see between all living organisms, ranging

<sup>2</sup>We leave RNA out for now which is very relevant for vaccine development.

from single cell up to the humans. The medical world can use that DNA to analyse the overall situation of the living organisms.

2. **Genome Sequencing - DNA & Chemistry to Digital** - The DNA will be read from the living organism but the total size of the DNA-map is very long. So the short reads need to be mapped on the overall DNA-string. This is done using classical computers and involves finding the location where the short-reads need to be placed.
3. **DNA Storage - Digital to Digital** - The result of the genome sequencing determines the genetic profile of the living organism. The genome of the living organism will stay the same, throughout the lifetime of that organism and thus needs to be stored. Currently, we use classical hardware and storage devices to do that.
4. **Genetic Analysis - Digital to Chemistry** - Based on that genetic profile, researchers from universities, pharmaceutical companies, and others, will further analyse the genetic information. Restricting our story line, this may lead to a diagnosis for a particular mental or physical sickness.
5. **Medical Diagnosis - Chemistry & Biomedical to Chemistry** - In view of the genetic analysis, pharmaceutical researchers will be developing a new medicine, vaccine or anything else to cure the patient.
6. **Medication Development - Chemistry to Biology** - After the diagnosis and the necessary research to develop an updated medicine, it will be developed and brought to the market.
7. **DNA of any Living Organism - Chemistry & Biomedical** The living organism, in our example the human being, will take the medicine and be cured.

This describes the entire process of a DNA-based organism, like the human, is determined by processing the entire DNA, stored in huge data centers up to the creation of a new medication to be used. We can continue and propagate this entire process at the level of all humans but also of all living organisms on this planet. When we limit our reasoning to humans, we already discover that we need a huge storage space to store all the DNA-information, which is stored in a binary way. As said in the introduction, many researchers have started looking at DNA to use it as the main way of storing information. Information needs to be interpreted in this paper as going beyond the DNA of living organisms. They are talking about any kind of digital piece of information that can be stored in the DNA-way. When looking at the entire process, we observe immediately that many different data formats are used to extract the relevant information from any DNA observation. But it is clear that we go from DNA in a chemical-biomedical structure, to digital, back to chemistry and ultimately in the chemical-biomedical generation of new medicines. It is exactly in the digital part that our planet is experiencing huge problems. A very interesting international collaborative project is **Q4Bio**<sup>3</sup>, in which also Wellcome participates. One of their goals is what I describe in this paper: assess quantum hardware when executing health applications, expressed in the QC-logic way and also using DNA-storage.

## 4 DNA Storage Principles

In this paper, we focus on DNA, not only for the pure medical or biological reasons, but also for the use of DNA as a completely new storage device. To have a meaningful and medical example, let us now simply assume the following. The medical world, in combination with the pharmaceutical companies, wants to develop medicine that are tailored to the specific genetic profile of the organism they are treating. That has two immediate consequences. Limiting now the discussion to humans, the **first** is that we need to have a genetic profile for every individual on this planet. That profile can be used to define the specific medication the person needs to receive. Generating a genetic profile, using machines from Illumina for Europe and North-America, is time consuming as there are not enough of those machines around. We do not have enough storage spaces to save all the genetic data of every individual on this planet, which is the long-term goal. The immediate but also long-term emerging need of this observation is that the genetic data for every individual human needs to be stored. That is how it can be used for any future medical analysis. The human genome needs around 1 GB of classical storage but when we scale that up to the 7 billion people living on this planet, we immediately see the limitations of this approach. The **second** direct consequence is to use DNA as a storage device. To store and manage the DNA for one living

---

<sup>3</sup>[wellcomeleap.org/q4bio/program/](http://wellcomeleap.org/q4bio/program/)

organism, but also DNA found from animals that died thousands of years ago, can still be reconstructed and analysed.<sup>4</sup> This observation led many researchers to look at DNA as a way to store huge amounts of information for up to extremely long times. The use of DNA can therefore also be used as a new way to store digital data as a biological molecule. We base ourselves primarily on [23] [21] [24] [25][26] [27][28].<sup>5</sup>

Based on the huge amounts of data we store in a digital way, there is a growing shortage of storage space. In addition, there is the rising number of scientific and industrial projects that will be launched world-wide which will generate zettabytes number of data that need to be stored and analysed.<sup>6</sup> We mention scientific and industrial projects without specifying a specific field or industry but the amount of data that will be collected for any kind of problem will be incredibly high. It should be clear that the situation of CMOS-transistors also has a serious effect for the storage of the data, in addition to the computing effect.

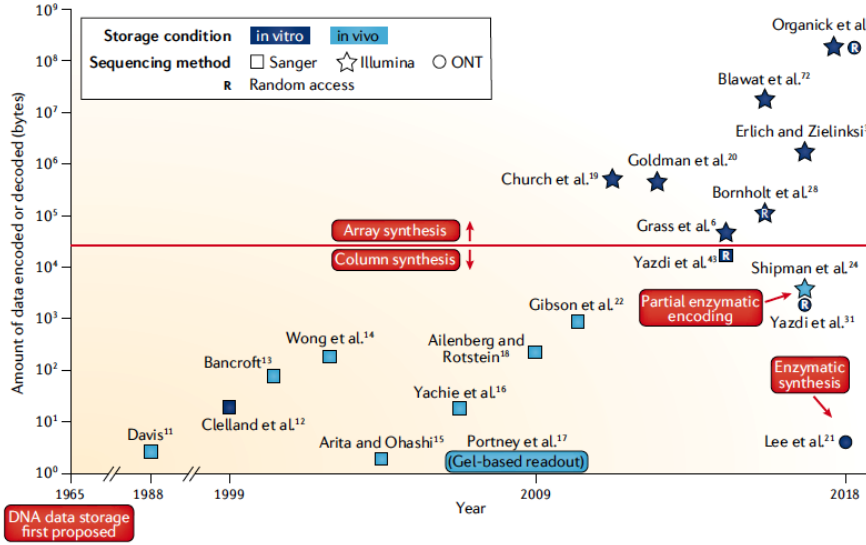


Figure 5: DNA Storage Research Overview [28]

DNA is the basis of any living organism, and is expressed in 4 amino acids, represented by adenine (A), cytosine (C), guanine (G), and thymine (T). They always combine in two characters, AT or TA and CG or GC and have a different base value than the binary way of representing data. [25]

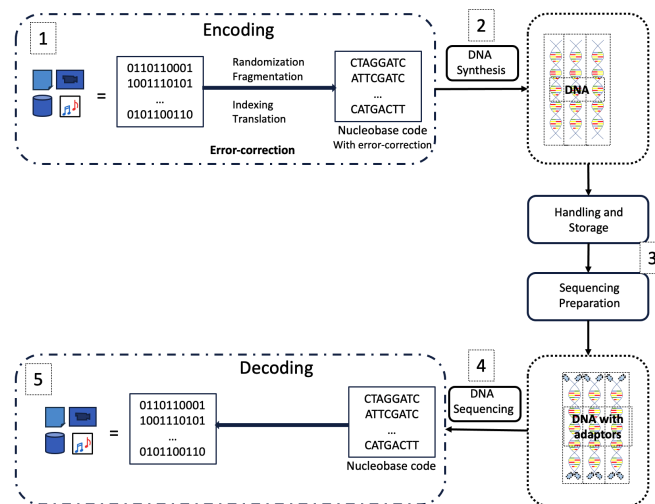


Figure 6: DNA site change from [25]

<sup>4</sup>We do not make any statement about bringing organisms back to life based on their DNA.

<sup>5</sup>Very interesting other papers are [21][24] [29]

<sup>6</sup>1 zettabyte= $10^{21}$  bytes

We give a short presentation of the protocol described in [23] [25] and more in detail in [26] which was developed to provide error-free storage by protecting the information using error-correcting codes. We extend where needed to make the link between classical data, DNA-storage and QC-Logic based computations.

## 5 The Full Stack for DNA-Data Storage

The **DNA Data Storage Alliance**<sup>7</sup>, established in 2020, comprises a diverse consortium of companies and academic institutions. Prominent industry players such as Illumina, Microsoft, and IBM, along with various universities and research centers, are pivotal members. Their involvement underscores the significant potential and industrial interest in DNA-based data storage solutions. The main goal is to use DNA as a new storage platform to solve all the storage and computing problems the classical hardware community cannot solve. This alliance clearly has as main objective to expand the classical way of storing data in data centers in a DNA-format and still use classical computers, and binary format, to process the data.

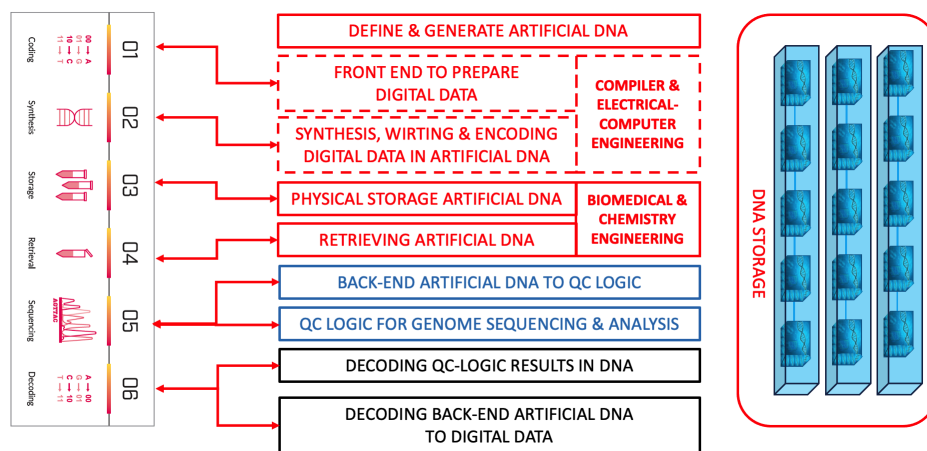


Figure 7: The DNA storage process from

Shown in Figure 7, and described in the first pages of report [23], downloadable from their website, observations are mentioned provided by, for instance, Gartner and IDC StorageSphere, where the amount of data that will be generated will grow between 19% up to 50% per year. The report mentions a long-term storage device called Linear-Tape-Open or LTO-tapes. Any DNA-device can store up to 115.000 more data than any LTO-tape.

Mentioned in that document, the following steps are defined and are part of the entire process to go from any data format such as digital, to storing it on the DNA-storage device, and at a later moment, to retrieve that information and generate the original digital item. We added and changed some of the layers.

1. **DEFINE & GENERATE ARTIFICIAL DNA** - An important aspect of any DNA-storing technology consists of creating, what is called, an **artificial DNA**. Some interesting papers are [30] [31] [32]. When going through those papers, one thing is quite evident. The authors all speak of medicine and health challenges to justify why developing an artificial DNA is needed. The first logical step is, therefore, to create an artificial DNA that will be used to translate all data elements to. This step precedes the execution of any further step in the full stack process.
2. **FRONT END TO PREPARE DIGITAL DATA** - Whatever piece of digital data needs to be encode to the artificial DNA needs to be prepared such that the translation can be done at high speed, by classical computers. It may involve creating multiple copies of the digital item such that multiple DNA-versions can be stored.
3. **SYNTHESIS AND ENCODING OF DIGITAL DATA IN ARTIFICIAL DNA** - In [23], encoding methods are presented that relate the synthesis and sequencing methods being used.

<sup>7</sup>dnastoragealliance.org

This way, bit density is enabled that ensures the reduction of error rates., enabling acceptable bit density that largely compensate for error rates. As mentioned in the previous step, this guarantees multiple copies of the original binary data, often by cutting the original files in smaller parts. **Error-correcting codes for encoding** are needed such that classical and binary data files for any existing object, being a movie, even an old music CD-ROM, a written text or an existing hard drive, can be translated in the artificial DNA-format. Error-correcting is done by adding redundant data to the original files, and then the updated binary data is translated to a set of sequence strings, composed of the four DNA bases. The encoding strategy starts from the binary data that will be mapped on the four bases of DNA, using the following:  $00 \rightarrow A$ ,  $01 \rightarrow C$ ,  $10 \rightarrow G$ ,  $11 \rightarrow T$ .

4. **PHYSICAL STORAGE OF DNA** - Using different ways, such as randomization, fragmentation, indexing and translation, we all translate the binary 0s and 1s on any of the four DNA-bases. There are two DNA-bases, the first one combines A and C, the second one uses G and T. Synthesis refers to the process where DNA is manufactured, involving various chemical steps. The DNA molecules are combined in various ways such that the bits-to-bases is mirrored. DNA Handling and Storage together with Sequencing Preparation are required to make the DNA readable for any DNA machine. It implies that the DNA is extended with predefined sequencing adaptors to both ends of the DNA molecules. Using these machines, all the new strands are sequenced such that they can be stored in, for instance, test tubes up to thousands of years. Important to underline is that, in this experiment, an Illumina machine is used which is very error-free. The amount of tolerated errors can be made larger, but that choice will result in fewer bits per nucleotide and result in more redundancy. Therefore, fewer result bits per nucleotide will increase the cost of storing data in DNA. After the synthesis, the DNA is enclosed and deposited in a library of large pools of DNA. This may assume the presence and use of inert gas or other chemicals, that guarantee their long-term preservation.
5. **RETRIEVING ARTIFICIAL DNA** - The digital data sometimes needs to be retrieved and sent back to the original data owner. It may involved making copies of the molecules for sequencing methods that are molecule intensive activities. It also allows to store multiple copies to serve distributed storage.
6. **SEQUENCING OR READING** - This is called the retrieval of DNA. Various sequencing methods are available, such as sequencing-by-synthesis (SBS) or nanopore sequencing. These are often optical, pH-based or electrical techniques such that the different strands of DNA can be read.
7. **BACK-END ARTIFICIAL DNA TO QC-LOGIC** - The main goal is to combine QC-logic versions of applications that can be executed on any future quantum computer device. For the rest of this paper, emphasis in QC-logic is the use of **perfect qubits**, that do not have any erroneous behaviour when performing any quantum gate on one or more qubits. In this paper, we will show the Quantum Genome Sequencing algorithm, which is crucially important for any short-reads we will get for any living DNA-organism. The goal is indeed that all data is stored in the artificial DNA, such that this step requires the translation of DNA-data in the quantum data format.
8. **QC-LOGIC FOR GENOME SEQUENCING & ANALYSIS** - The current version of QG sequencing and analysis uses the quantum data to achieve a result, which can then be encoded again in the artificial DNA that we used earlier in the process.
9. **DECODING BACK-END ARTIFICIAL DNA TO DIGITAL DATA** - The final step is to translate the data stored in the artificial DNA back into the original digital data items. This is also the step where any error correction takes place as any of the previous steps may have generated some problems. The last step of the decoding process is to reassemble the data in the original digital form, and make it available for the end-user. The goal of any DNA-storage is to reuse the same data at some point in the future. We have to reconstruct the initial binary data by translating the DNA-code again to a binary code. As the use of the decoded DNA strands that are converted back to their original binary format is much simpler than the DNA generation process, this step mostly has to take possible errors into account. Depending on the error-requirements, one can have high or very low error frequencies. Several ways of dealing with those two situations are described in [33], indicating that it is impossible to avoid errors and that the higher the error frequency is, the higher the cost to retrieve data as more data needs to be retrieved and tested.



## 6 Advantages and disadvantages of DNA-storage

As enumerated in several papers such as [23] [34] [21], the advantages of DNA are the following.

1. **Durability and Stable over time** - DNA can be preserved for thousands of years under minimal conditions, such as being stored in a cellar, making it an exceptionally long-lasting storage medium.
2. **Density and Storage Capacity** - DNA has a really big storage capacity. DNA with a weight of half the weight of a sugar cube can store up to 215 million GB of data, for a very long time. Another example is that around 450 Exabytes can be stored in a DNA-device the size of a shoe box.
3. **Format Immutability** - Better predictability is expected through the use of complementary nucleotides. These nucleotides are always the complementary combinations of the amino acids which make the synthesis and sequencing easier due to the predictable pairing of A with T and C with G.
4. **Energy Efficiency and Sustainability** - DNA does not become obsolete and avoids the environmental impact associated with conventional storage materials. Very few energy is needed to store and even transport the DNA-data. Copying DNA-data should also lead to few errors that can be corrected.
5. **Cost Effectiveness** - Synthesising DNA is relatively inexpensive and is becoming more accessible as technology progresses.
6. **Parallel computing** - An unexpected advantage of DNA-computing is the way that DNA-encoded information is executed in parallel. Parallel programming on a classical computer

There are also a number of serious disadvantages related to DNA storage. In [35], the authors separate between **write-once-read-never** with an access frequency of once every 10 years, a **working storage** where the DNA is accessed multiple times per year, and finally the **short-term storage for dynamic handling of data**.

### Long-term Storage

First the **write-once-read-never** approach.

- **Potential contamination of DNA** - When storing DNA for a very long time, there can be contamination from contemporary bacteria or even human DNA. The oldest DNA is estimated to be 400.000 years old and it is difficult to verify how the original DNA properties were.
- **Optimistic estimations of DNA stability** - When basing ourselves on recovered DNA from fossils, certain genetic DNA properties are difficult to isolate.
- **Short strands of DNA** - It is necessary to develop synthesis technologies that can make longer strands. That way the percentage of overhead per strand can be reduced, hence increasing the information density of the system.

### Working Storage

When DNA is used as a **working storage**, with accesses limited to multiple times per year, the following needs to be taken into account.

- **Full recovery when stored less than 2 years** - When the storage is less than 2 years, there is an almost perfect recovery possible, either when frozen or in aqueous solution or even as a dry solid. There are experimental results that storage at 4 degrees C or below in aqueous or dried form can provide stability for a couple of years.
- **Degradation due to information access** - Each time DNA needs to be accessed, it may lead to rehydration of dried DNA or other effects.
- **Freezing leads to more breakage** - There are indications that frozen DNA have a bigger degradation when accessed. Observation results are found to have up to 75% of degradation after 20 freeze thaws.

## Short-term Storage

Finally, the **short-term storage** for dynamic handling of data also leads to a couple of serious challenges.

- **Error-tolerant encoding of DNA** - Depending on the way the DNA is stored and how frequently it is used, it may be important to have an efficient and fast error-tolerant encoding system installed.
- **Tuning DNA stability** - We need to understand the parameters to guarantee good stability of the DNA when using it as a storage approach.
- **Physical manipulations will create damage** - It will be very difficult to avoid damage when doing any kind of physical manipulation of the DNA. This is certainly part of the research agenda one needs to define.
- **Operational temperatures** - Sometimes the temperatures for any manipulation of DNA should be low enough to reduce the degradation of the DNA. The presence of phosphate and high temperatures are seen as critical elements.

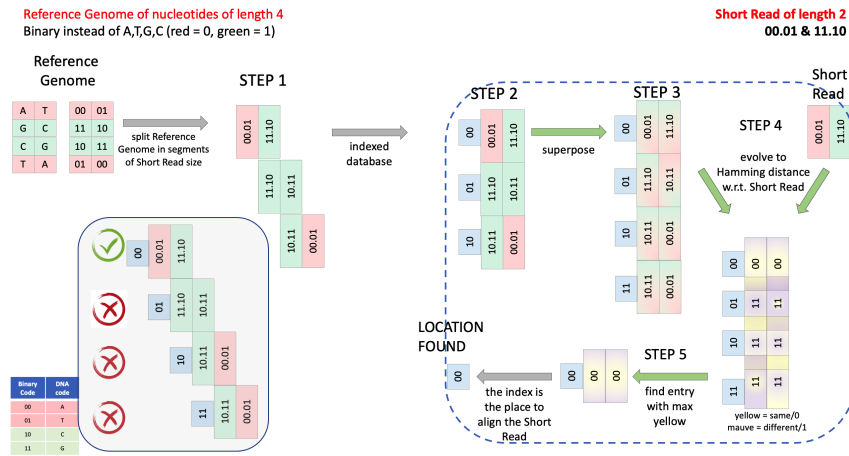


Figure 8: Quantum Genome Sequencing Algorithms

## 7 Combining Quantum Genome Sequencing and DNA-storage

This section in the paper relates the quantum computing challenges with the data storage capacity that DNA has already demonstrated. The goal now is to illustrate that there are very large but, for now, theoretical compute possibilities based on qubit technology. We have to assume that it will take another 10-15 years to be realised. There is however a second huge challenge for humanity on this planet, which is to store all the accumulated data we have now, and the amounts that will only increase in the coming years, in data centers. The combination of compute power and data storage are the two assumed capabilities for the two technologies that, together, will build a new computing platform. It suffices to state that quantum computing has a potential to perform all kinds of computations in an implicit parallel way. No parallel version of any algorithm needs to be written by humans, but the quantum accelerator will generate all the possible paths in the parallel version of the algorithm and all the parameters to access all paths.

We illustrate in this part the essence of genome sequencing and conclude this section by introducing how they could be combined. Without going in the details of the quantum version of the genome sequencing algorithm, as can be found in [36], we sketch the broad lines of the quantum version of the genome sequencing algorithm. We emphasise that we only adopt here one particular approach, even though many others are also available and have been tested by us.

Based on Figure 8, we can also see that a Genome Sequencing algorithm is based on the binary format of the initial DNA-reading, performed by a DNA-machine. They need to be represented in a Quantum Computing logic way, abbreviated as **QC-logic**, to start executing the algorithm. Important to realise is that inside the dotted line of Figure 8 all operations follow the QC-logic way of thinking. The entire process, from DNA, QC-logic and the use of digital data is described.

- **Reference Genome** - Any DNA generated via a DNA-machine needs to be mapped on the reference genome, which is common for all human beings on this planet. Evidently, we cannot model the logic for the entire human genome, implying that the figure represents a substantially reduced reference genome.
- **Short Read** - The short read, generated by the DNA-machine, is shown on the right of the figure. This short read needs to be mapped to the reference genome. Any DNA-machine will generate all the short-reads that need to be mapped on the reference genome.
- **Step 1 - DNA Short Reads** - This step represents the sequence of short reads that need to be mapped on the reference genome. A serial representation is made of the short reads.
- **Step 1 - Full Short Read List** - Next, the series of short reads is generated, that will be mapped on the reference genome. It can be seen that the short reads all map on the reference genome, which is the ideal case.
- **Step 2 - Indexed Short Read List** - This step is the first QC-logic step, inside the dotted line of the figure. All the short reads are now indexed.
- **Step 3 - Superposition** - Based on the short reads, all the components of the reference genome are placed in a specific quantum way, called the superposition. This guarantees that all the possible combinations are included in the reference genome map. That particular set of genomes will be used to identify where the short read maps in the reference genome.
- **Step 4 - Hamming Distance** - The next step is to see where the short read maps in the reference genome. In our QGS algorithm, the Hamming distance is used. When a cell is coloured 'yellow', that means there is a match, and, in our example, there are two short read cells. That implies that the other cell of the short reads also needs to be verified. Two yellow cells represents a perfect match. One or two cells can be grey, which shows a partial or no match.
- **Step 5 - Location Found** - In our example, the short read maps on the first genetic pair in the reference genome.
- **Index of Short Read Location** - Ultimately, the location is communicated to the final list of short reads that are used to find the exact location of the current short read.

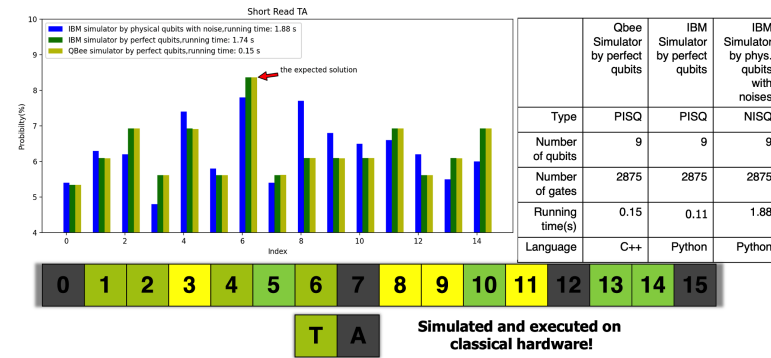


Figure 9: QC results on QBeeSim and Qiskit

Figure 9 shows the graph that our own simulator, called QBeeSim, generates, together with the graph generated by the IBM-platform Qiskit. What we emphasise here is that both IBM as well as our simulator allow quantum algorithms to be formulated in **perfect qubits**. That means that they do not decohere nor have any errors in the quantum gates used. It is very interesting that also IBM supports such a concept in the world-wide used Qiskit platform. What is important to notice is that, when using classical supercomputers to simulate quantum circuits, one can go up to around 70 qubits in superposition before the entire memory of the supercomputer is consumed. There is a critical need for substantial investment in QC-logic across various scientific fields to cultivate experts capable of developing and implementing QC-logic based algorithms.

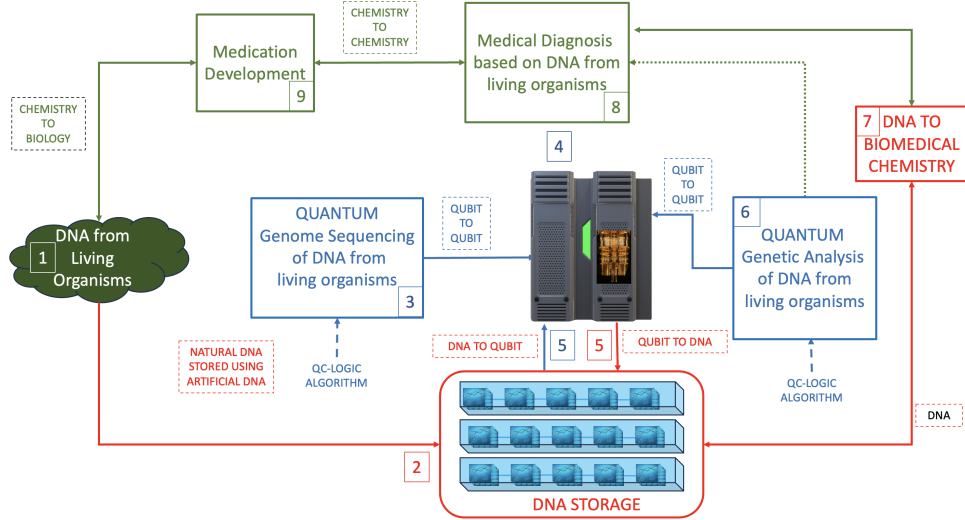


Figure 10: Quantum medical use of DNA Genome Sequencing and Analysis with DNA Storage

In parallel, it is equally important to be able to interpret and apply DNA technologies effectively. We should also not forget that the entire QGS algorithm always assumes a digital representation of the DNA short reads. The goal of using DNA-storage devices will change that completely to a DNA-representation. The QC-logic steps will stay the same, but the start and end of the entire GS process will not be digital.

Figure 10 is a concrete example of how any data set can be stored in a DNA-storage device. We can also assume that such a storage device is connected to a quantum computer, providing data on which a quantum algorithm can be executed. Evidently, it could also be a classical computer but this paper describes the future of computing and storage platforms.

Given the context of this paper, we give the example of genome sequencing, developed and written in a QC-logic way, and the use of a DNA-store. It is an example of how medical genome sequencing and analysis data is being generated from any living organism, but it also shows that the genetic data, extracted from the living organism, is stored in a DNA-storage device. This is the theoretical approach and is not yet developed. It is clear that we combine two approaches. First, the quantum version of genome sequencing and analysis is executed on a future quantum computer. Second, a DNA-storage device is used to store the huge amounts of data are needed for the genome sequencing and analysis.

Figure 10 shows the different steps in the process of quantum genome sequencing and analysis will need. The starting point, called **Step 1**, is to read the DNA of any living organism, such as a human. **Step 2** is to store that DNA in a DNA-storage device, ready to be used by the QGS-algorithm. The next step is to send the quantum algorithm, **Step 3** to the quantum accelerator, **Step 4**. When executing the QGS-algorithm, the accelerator will need the original DNA which was obtained in Step 1. To this purpose, the DNA-data which needs to be translated to qubits, **Step 5**. A similar step can be done for the genetic analysis if that is what the user desires to execute **Step 6**. Again, the DNA-data needs to be translated to a quantum-data format. The results of the quantum algorithm will be stored again in the DNA-storage device, implying a translation from the quantum-format to the DNA-format, **Step 5**. The final results of the analysis of any living organism DNA will be stored in the DNA-storage. That information needs to be transferred to the Medical Diagnosis unit that will look at any pharmaceutical or medical solution that can be implemented. Represented by **Step 7**, this involves the translation or connection with biomedical chemistry and sending it to the diagnostic unit, **Step 8**. Based on the diagnostic reflection, the medical solution will be development by the Medical Development unit, **Step 9**, ultimately resulting in a specific treatment or medicine for the living organism, the human in our case. Important to note here is that only the different steps are described, and we do not mention the development of new devices or any other medical tests that need to be performed before any medical treatment can be made.

A first intermediate conclusion can be formulated. As explained before, any kind of digital data can be stored in a DNA-format such as movies, paintings, books, photographs etc. Not only the energy is substantially lower as compared to storing data on classical hardware, but, more importantly, the volume needed to store huge amounts of digital data is so low. In [34], the authors mentioned that petabytes of data can be stored in a DNA-device the size of a shoe-box. Remember that, currently, we need data stores

of several hundreds of cube meters and cooled to temperatures between 16 and 20 °C. It is clear that this DNA-direction to store the huge amounts of digital data needs to be explored and further researched.

## 8 The New Path Forward

In order to arrive at a global conclusion to be discussed world-wide, between researchers from industry, universities etc, we present Figure 11. Currently, the world-wide computer infrastructure is fully based on classical hardware and software. This Figure gives a world-wide use of the technologies we are proposing, and on which many people have been working. The contribution of this paper is to combine those two, hopefully contributing to address the fundamental technology-driven challenges humanity has to solve.

We add both quantum computing hardware and software but also the DNA-based alternative for data-storage. It is very clear that the quantum computing R&D effort is understandably dominated by the physicists, mathematicians and electronic engineers. However, we argue that researchers from all scientific fields, ranging from computer science to space and chemical engineering need to adapt an important and fundamental direction in their research and development. This will have a substantial impact on all the scientific and even economic efforts world-wide. We therefore ask all scientists to **explicitly** and vigorously embrace the main message of this paper, and that focus on both computing as well as storage of data.

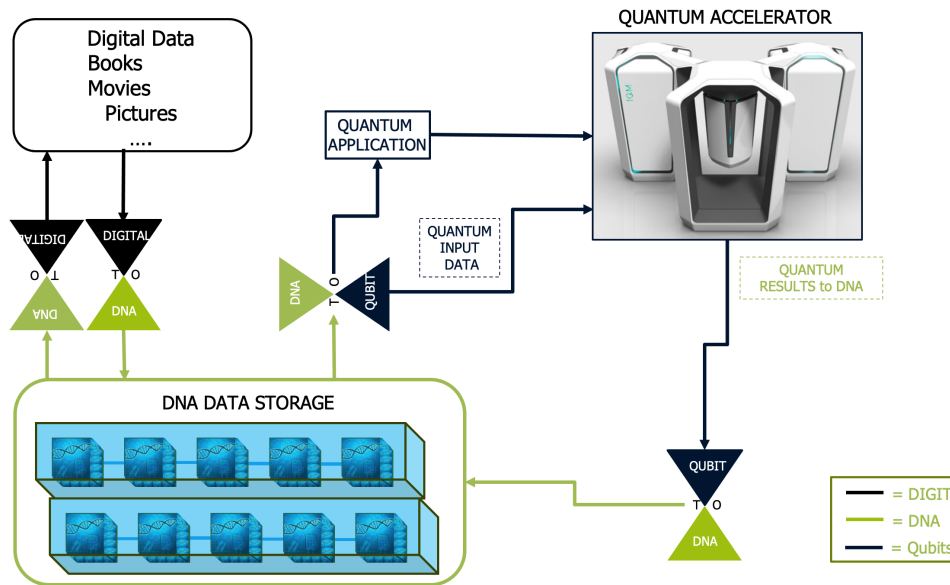


Figure 11: Future computing platform, combining QC and DNA-storage

Regarding the quantum computing part, we introduced the the PISQ-oriented research line. **PISQ** stands for **P**erfect **I**ntermediate-**S**cale **Q**uantum computing and is a complementary approach to **NISQ**. The **PISQ** approach provides an abstraction layer, so no direct reliance on the roadblocks and/or progress of the quantum hardware-based chip development efforts. As explained in [2] [3], the ultimate goal is to reach the Fault-Tolerant Quantum Computing level where all the physical problems of quantum computing have been solved. Certain industrial fields are already looking at the quantum implications for their domain. We have started working on Earth Observation, quantum genetics and quantum computational chemistry. Especially in the last two fields, we have already achieved good results but a lot of international collaboration is still needed. This is where industrial companies and universities can collaborate much more. The quantum hardware physicists working on any qubit technology will have to find ways to solve all the known challenges. This will also have repercussions on the kind of micro-architecture that is needed. When developing new quantum gates, the quality of the qubits will certainly improve. This is what the evolution of the error rates represent. Hopefully, the scalability will also get better such that we can use more qubits in the quantum applications. As stated, small research groups, such as the one from prof. Morello at UNSW in Sydney, are among the best in the world on one particular kind of qubit technology and only focusing on small number of high-quality qubits. Their results could be bought by some of the large players, who are focusing on the large number of qubits. With all the errors they still have.

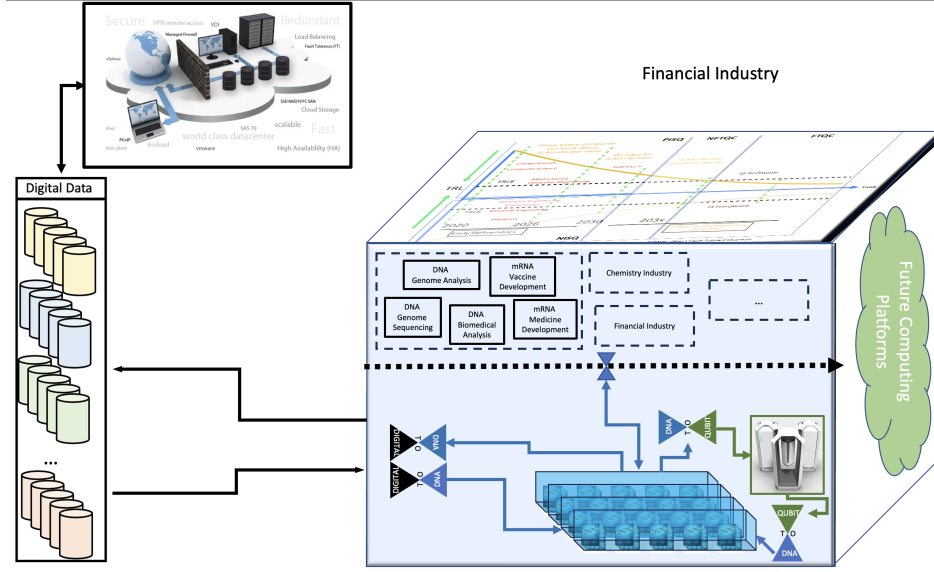


Figure 12: Long term infrastructure: Classical, Quantum Computing and NDA-storage

The final figure in this paper, Figure 12, shows how any future computing platform will be combined. Evidently, classical computers are so present in all aspects of our daily lives, that they will continue to be present. It is clear that any DNA-storage device also will be a crucial part of any professional or personal life on this planet. Evidently, we will also need extremely powerful and new supercomputers. Those kinds of computers will be implemented using any one or two kinds of quantum technology choices will mature.

The story line as described in this paper clearly shows that in terms of pure storage capacity no other technology than DNA-based is capable of storing enormous amounts of data, already available or being generated in the coming years.

## 9 Conclusion

The final conclusion of this paper, shown in Figure 12, is that QC and DNA-storage are very complementary approaches for the large problems of this planet, namely to analyse the huge amounts of data that humanity has already assembled. It is absolutely certain that, in about 10 years, classical CMOS-based hardware will stop growing in compute power. On top of that technological constraint, there is an additional problem of storing the data that we are continuously collecting. The quite recent creation of the **DNA Data Storage Alliance** is a very important indication that large companies are focusing on that particular challenge.

The message is that we should develop both the compute as well as the storage objectives. A logical way would be to develop both objectives in parallel, but, as we suggest, to connect them as soon as possible. Our starting point is the use of the quantum genome sequencing algorithm. Rather than basing it on a digital representation of the DNA short-reads, we want to feed the quantum accelerator with the DNA-data translated to QC-logic terms. This will guarantee to address different goals such as scalability and speed of the DNA-storage technology as well as the QC hardware and software.

We already mentioned that we are entering a completely new way of doing research and development. Other fields need to get involved, in terms of data storage and computing. The software part should also be initiated in parallel with the hardware development. Software can be defined as a hybrid way of developing applications, where binary, quantum and DNA-items are combined.

As explained here, synthesizing DNA, along with its storage and retrieval, remains a slow and costly process, with the cost of writing one petabyte of DNA estimated at one trillion dollars. Enzymatic DNA Synthesis presents a promising improvement, using enzymes to add bases to nucleotide chains more efficiently. To further advance DNA data storage technology, the focus should be on developing new DNA printing methods, increasing the length of DNA chains to enhance capacity and efficiency, and reducing synthesis costs to make the technology economically viable.

At the quantum computing level, we need to focus much more on developing quantum algorithms for

many of the large problems, where chemistry, medicine and space are immediate candidates on which international collaboration is possible. It is very difficult to predict what and how technology will evolve to build a scalable and robust computing platform for the future putting the emphasis on intensive collaboration world-wide.

## References

- [1] Melvin M. Vopson. The information catastrophe. *AIP Advances* 10, 2020.
- [2] Koen Bertels, Aritra Sarkar, and Imran Ashraf. Quantum computing—from nisc to pisc. *IEEE Micro*, 41(5):24–32, 2021.
- [3] Koen Bertels, Emma Turki, Tamara Sarac, Aritra Sarkar, and Imran Ashraf. Quantum computing – a new scientific revolution in the making, 2024.
- [4] Sarkar A. *Applications of Quantum Computation and Algorithmic Information: for Causal Modeling in Genomics and Reinforcement Learning*. PhD thesis, Delft University of Technology, 2022.
- [5] Aritra Sarkar, Zaid Al-Ars, and Koen Bertels. Quaser: Quantum accelerated de novo DNasequence reconstruction. *Plos one*, 16(4):e0249850, 2021.
- [6] Aritra Sarkar, Zaid Al-Ars, Carmen G Almudever, and Koen Bertels. An algorithm for DNA-read alignment on quantum accelerators. *arXiv preprint arXiv:1909.05563*, 2019.
- [7] Ray KS. Mondal M. Review on dna cryptography. *Int J Bioinfor Intell Comput.*2(1), pages 44–72, 2023.
- [8] Hongfeng Zhang, Aritra Sarkar, and Koen Bertels. A resource-efficient variational quantum algorithm for mrna codon optimization, 2024.
- [9] M. Troyer T. Hoeffler, T. Haner. Disentangling hype from practicality: On realistically achieving quantum advantage. *Comm. of the ACM*, vol. 66, pages 82–87, 2023.
- [10] Terhal B. et al. Single quantum querying of a database. *Arxiv*, pages 1–6, 2009.
- [11] G.L. Long. Searching an unsorted database in quantum computers and duality quantum computers. In *Fifth International Conference on Natural Computation*, 2009.
- [12] Israa Hamouda, Ayman M. Bahaa-Eldin, and Hazem Said. Quantum databases: Trends and challenges. In *2016 11th International Conference on Computer Engineering & Systems (ICCES)*, pages 275–280, 2016.
- [13] Israa Hamouda, Ayman M. Bahaa-Eldin, and Hazem Said. A generalized grover’s algorithm with access control to quantum databases. In *2016 11th International Conference on Computer Engineering and Systems*, pages 281–285, 2016.
- [14] U. Çalikyılmaz et al. Opportunities for quantum acceleration of databases: Optimization of queries and transaction schedules. In *Proceedings of the VLDB Endowment*, volume 16, 2023.
- [15] Umut Çalikyılmaz, Sven Groppe, Jinghua Groppe, Tobias Winker, Stefan Prestel, Farida Shagieva, Daanish Arya, Florian Preis, and Le Gruenwald. Opportunities for quantum acceleration of databases: Optimization of queries and transaction schedules. *Proc. VLDB Endow.*, 16(9):2344–2353, may 2023.
- [16] Xie Shi-man and Shang Xin-zhi. The building and optimization of quantum database. *Physics Procedia*, 25:1602–1609, 2012. International Conference on Solid State Devices and Materials Science, April 1-2, 2012, Macao.
- [17] Manisha J Nene Sachin Khurana. Implementation of database search with quantum computing: Grover’s algorithm vs linear search. In *International Conference on Ambient Intelligence, Knowledge Informatics and Industrial Electronics (AIKIE)*, 2023.
- [18] Mikhail Dyakonov. When will useful quantum computers be constructed? not in the foreseeable future, this physicist argues. here’s why: The case against: Quantum computing. *IEEE Spectrum*, 56(3):24–29, 2019.



- [19] Leonard M. Adleman. Molecular computation of solutions to combinatorial problems. *Science*, 266(11):1021–1024, nov 1994.
- [20] Hui L. et al. DNA-based programmable gate arrays for general purpose DNA computing. *Nature* 622, pages 292–300, 2023.
- [21] Seok Joo Kim, Woo-Bin Jung, Han Sae Jung, Min-Hyun Lee, Jinseong Heo, Adrian Horgan, Xavier Godron, and Donhee Ham. The bottom of the memory hierarchy: Semiconductor and DNA data storage. *MRS Bulletin*, 48(5):547–559, May 2023.
- [22] Andrea Doricchi, Casey M. Platnich, Andreas Gimpel, Friederikee Horn, Max Earle, German Lanzavecchia, Aitziber L. Cortajarena, Luis M. Liz-Marzán, Na Liu, Reinhard Heckel, Robert N. Grass, Roman Krahne, Ulrich F. Keyser, and Denis Garoli. Emerging approaches to DNA data storage: Challenges and prospects. *ACS Nano*, 16(11):17552–17571, 2022. PMID: 36256971.
- [23] DNA Data Storage Alliance. An introduction to DNA data storage, 2021.
- [24] George M. Church, Yuan Gao, and Sriram Kosuri. Next-generation digital information storage in dna. *Science*, 337(6102):1628–1628, 2012.
- [25] Linda C. Meiser et al. Synthetic DNA applications in information technology. *Nature communications* 13, 2022.
- [26] Koch J. et al Meiser L.C., Antkowiak P.L. Reading and writing digital data in DNA. *Nat Protoc* 15, pages 86–101, 2020.
- [27] Joao Gervasio, Henrique Oliveira, Andre Martins, Joao Pesquero, Bruno Verona, and Natalia Cerize. How close are we to storing data in DNA? *Trends in Biotechnology*, 42, 09 2023.
- [28] Strauss K. Ceze L., Nivala J. Molecular digital data storage using DNA. *Nat Rev Genet* 20, pages 456–466, 2019.
- [29] Goldman N. et al. Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. *Nature*, 494, 2013.
- [30] Silver P. Smolke C. Informing biological design by integration of systems and synthetic biology. *Cell* 144, pages 855–9, 2011.
- [31] Hao N et al. Tunable signal processing through modular control of transcription factor translocation. *Science*, pages 460–4, 2013.
- [32] Tan C. Ding Y, Wu F. Synthetic biology: A bridge between artificial and natural cells. *Life* 4, pages 1092–116, 2014.
- [33] Organick L. et al. Random access in large-scale DNA data storage. *Nat Biotechnol* 36, page 242–248, 2018.
- [34] Zielinski D Erlich Y. DNA fountain enables a robust and efficient storage architecture. *Science*, 355:950–954, 2017.
- [35] Keung A.J. Matange K., Tuck J.M. DNA stability: a central design consideration for DNA data storage systems. *Nat Commun* 12, 1358, 2021.
- [36] Aritra Sarkar. Quantum algorithms for pattern-matching in genomic sequences. Master’s thesis, Delft University of Technology, June 2018.
- [37] J.W. Byers et al. A digital fountain approach to reliable distribution of bulk data. *ACM SIGCOMM Computer Communication Review*, volume 28, 1998.
- [38] M. Luby A. Shokrollahi. Raptor codes", foundations and trends. *Communications and Information Theory*, 6:213–322, 2011.



## A The DNA Fountain

A quite recent and very strong approach was developed by Erlich and Zielinski. [34] Their results show that they approach the Shannon capacity and guarantee robustness against data corruption.<sup>8</sup> The Fountain encoder works in multiple steps, as shown in Figure 13 and is based on oligonucleotide, called “oligo”, synthesis. The goal of storing the DNA is not on paper or any other storage device but rather using a DNA-approach. The authors use the nucleotide, where sugar can be combined with phosphate, sugar and phosphate etc. We will briefly describe the three steps in the algorithm, as described in the supplement of the Nature paper.

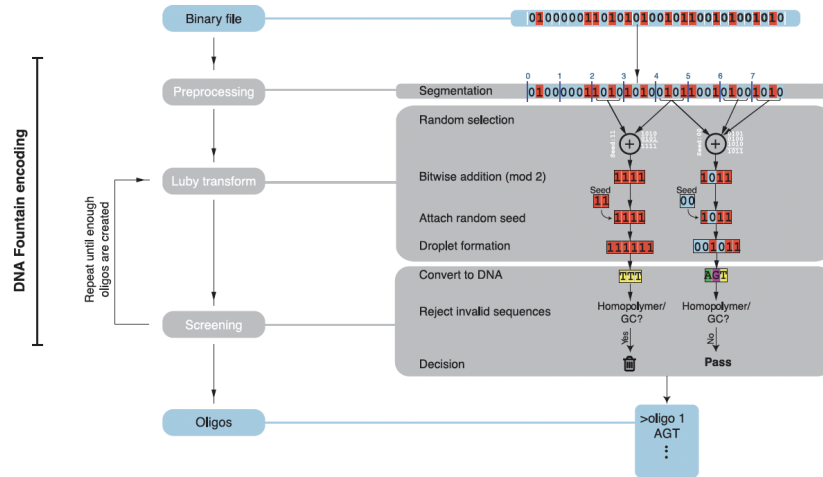


Figure 13: The DNA Fountain

### Step 1 - Preprocessing

Any digital storage item can be encoded in the DNA-code. For example, 00 could become the A, 01 translates to T, 10 to G and 11 to C. The human genome is 3.2 billion bases of AT or GC long. 14 million bases would fit on a small disk, shorter than the point at the end of a sentence. The original authors of this storage approach claimed that they could store 215 petaBytes of data in a single gram of DNA. The storage capacity of DNA is computed at one zettabyte in one gram of DNA. One zettabyte is 1 trillion gigabytes of data.

In the pre-processing phase, a binary file is pre-processed in a series of non overlapping segments of a certain length, using a standard lossless algorithms such as Gzip. This compression reduces the size of the input file, but it also reduces local correlations which are important for the screening step. The authors used 256 bits or 32 bytes for the **length L**, which is the number of bits. This results in a compressed file with non-overlapping segments of length L.

### Step 2 - Luby Transform

The second applies the Luby Transform to the file in the following way. [37][38] It first uses a pseudorandom number generator (PRNG) with a seed, selected from a mathematical rule. The algorithm then decides on **d**, the number of segments to package in the droplet. This number d is selected from the **robust soliton probability distribution**, shown in Figure 14. That particular distribution<sup>9</sup> is an efficient way to construct a good degree of encoding symbols. It will ensure that most of the created droplets, represented by **K**, have a small number of input segments. Once the parameter **d** is chosen, the PRNG-method will select d segments without replacement from the total number of segments. Then, the algorithm performs a bitwise-XOR operation, where two bit patterns of equal length are used. The result is **1** if only one of the bits is 1. The result will be **0** if both bits are 0 or both are 1. For example, use the droplet 0100,1100,1001, which will give **0001** as a result.<sup>10</sup> The authors state that the fixed-length index determines the binary representation of the seed. As the seed could be equal to 3 bits, and 2 bits

<sup>8</sup>A good TEDx talk by Zielinski can be found here: <https://www.youtube.com/watch?v=wxStLzuxxCw>

<sup>9</sup>Very good explanation by Lunny: [www.youtube.com/watch?v=C4qi\\_oJoUrE](http://www.youtube.com/watch?v=C4qi_oJoUrE)

<sup>10</sup>0100  $\oplus$  1100  $\oplus$  1001=0001

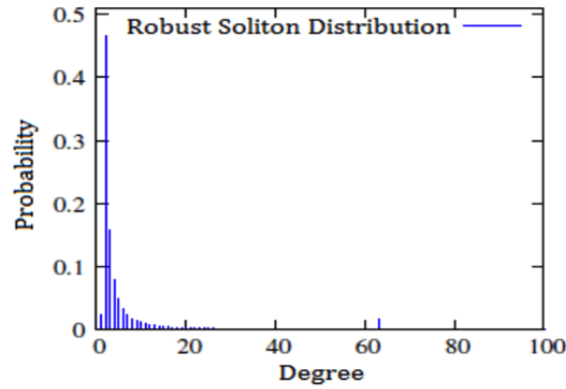


Figure 14: The Robust Soliton Distribution

represent the fixed index length, the output droplet will be **1100001**. The Luby Transform will prioritise robustness over dropouts by producing more oligos than the dropout rates. That also allows to recover the file as long as one has more droplets than **K**. The term fountain or rateless refers to the fact that these codes do not exhibit a fixed code rate.<sup>11</sup> Fountain codes allow data to have a any sequence size of encoding symbols such that the original symbols can be recovered from any subset.

### Step 3 - Screening

The final step is the screening of the droplets. All droplets that violate the required biochemical constrains of the DNA-sequence. Going back to our first example, we can assume that **00,01,10,11** can be translated to **A,C,G,T**. The example proposed earlier has **110001** as droplet that is translated to **TAC**. The algorithm then screens that sequence according to the DNA-properties. If the screen is positive, the droplet is considered valid and is added to the oligo design file. Important to emphasise is that the number of droplets in the valid oligo follows the **robust soliton distribution**.

### Step 4 - Oligo generation

The final step in the encoding phase is the oligonucleotide, called “oligo”, synthesis. The output is again a fully consistent DNA sequence that can be stored chemically in the test tube. From that chemical substance, the original data can be recontructred using the decoding approach, which we will not present in this paper.

<sup>11</sup>[https://en.wikipedia.org/wiki/Fountain\\_code](https://en.wikipedia.org/wiki/Fountain_code)