DNA Storage: Synthesis and Sequencing Semiconductor Technologies

Dominique Lavenier

DNA Storage: Synthesis and Sequencing Semiconductor Technologies

Dominique Lavenier
University of Rennes, IRISA-CNRS, Inria, Rennes, France

Abstract – Storing data on DNA molecules is an alternative to traditional storage media by promoting density and longevity. This paper gives an overview of how data is stored on DNA and how the reading and writing processes (sequencing and synthesis) are currently implemented on semiconductor devices.

I. INTRODUCTION

IDC (International Data Corporation) estimates that the growth rate of data between 2020 and 2025 is about 23% per year. By 2025 the volume of data generated is expected to reach 175 Zetta Bytes, of which nearly 12% (22 ZB) will need to be stored [1][2]. Most of these data (80%) will be hosted on large public or private data centers, and stored on SSD or HDD devices.

When considering the lifetime of data, it is interesting to note that the probability of a specific piece of data being accessible after 90 days becomes very low. In fact, 60% of these data become "archived" after this period. In addition, many data must be stored for a very long time for medical, legal or other reasons and will never be accessed again. You simply need to have these data available in case you need them. These data are called "cold data". Although they require no further processing, the cost of storing them is not zero since the electronic storage devices (SSDs, HDDs, tapes) housed in a data center must be replaced periodically.

In this context, DNA storage appears to be a promising alternative for this category of data. Indeed, two main criteria militate in favor of this technology: density and longevity.

Density: DNA is a very compact polymer. To give an idea: each human cell contains 6.4 billion nucleotides (A, C, G or T), spread over 23 pairs of chromosomes, and representing approximately 1.6 GBytes of information. Knowing that a human body is composed of 3×10^{13} cells, its storage capacity would be 48 Zetta Bytes of data! And the DNA molecules themselves represent only a very small fraction of the composition of the cells.

Longevity: DNA molecules can remain intact for thousands of year at room temperature in a dry atmosphere. The extraction of DNA from ancient fossils is proof of this.

In addition to these two main characteristics, we can also mention the stability over time of this storage medium: as long as humans exist, there will always be DNA readers and they will always be faster, smaller, more reliable, and cheaper. An important point is also the TCO (Total Cost of Ownership) since once DNA polymers are created, there is no need to periodically do anything else.

II. DNA STORAGE PRINCIPLE

In what follows, DNA molecules are used as simple polymers independently of any biological purpose. They are not inserted into a living organism to perform read/write, duplicate or other operations. DNA polymers are just used as an information carrier allowing the linking of 4 nucleotides represented by the four letters A, C, G and T (Adenine, Cytosine, Guanine, Thymine). The storage of information on DNA consists in assembling chains of nucleotides whose sequence reflects the stored information. Let us recall that a DNA molecule is composed of two complementary strands with the following pairings: A-T and C-G. Thus, from a single strand (template), a DNA molecule can be completely synthesized.

The entire DNA storage pipeline can be divided into six different steps, as shown in Figure 1. The first two steps correspond to writing the data. The last three steps correspond to reading the data..

1. **Encoding:** This step takes as input binary documents (text, image, music, video, etc.) and translates them into A,C,G,T sequences. Because of the errors generated by the following steps, specific DNA encoding is required, as well as powerful error correcting codes. In addition, sequence indexing is necessary because many documents will be physically stored in the same storage space.

2. **Synthesis**: Each A,C,G,T sequence is transformed into one or more real DNA molecules. The synthesis of a sequence of N nucleotides requires N cycles, each cycle being divided into several chemical reactions to sequentially design the molecule, nucleotide by nucleotide. The synthesis of a sequence generates a pool of identical DNA molecules.

3. **Storage:** DNA molecules are usually dehydrated and encapsulated into small containers in a controlled atmosphere suitable for long-term storage. The Imagene[1] company, for instance, offers a room temperature solution for the storage of DNA polymers [3].

4. **Retrieval:** A container may contain millions of different documents, each linked to many DNA molecules. To extract a single document, a PCR based on

[1] http://www.imagene.eu/

the sequence indexing scheme can be performed to select only the desired DNA molecules, i.e. molecules attached to one document.

5. **Sequencing:** DNA molecules are translated back into A,C,G,T sequences using DNA sequencers that are currently used for genomic analysis.

6. **Decoding:** The ACGT sequences generated by the previous step are first corrected before being reorganized to reconstruct the original document.

The critical part of this process is the synthesis stage. First, it is a slow process that today does not yet reach the speed required for DNA storage to be a competitive alternative to other storage technologies. Second, the cost of synthesis is too high to make this technology economically viable at this time.

The sequencing step, in comparison, does not represent a bottleneck. The considerable progress made in this field over the last 20 years has allowed for a relatively rapid reading of DNA molecules. Moreover, the ever-increasing needs in genomics are pushing these technologies to be ever more efficient.

The two next sections explains in more detail how to write and read information on DNA, and how solid-state devices are used to perform these operations.

III. SYNTHESIS

Currently, DNA polymer synthesis is limited to short, single-stranded oligonucleotides (length < 200 nucleotides), abbreviated in the following as "oligo." Phosphoramidite chemistry is mainly used to synthesize oligos. It consists in sequentially adding one nucleotide after another. The addition of a nucleotide is achieved by a succession of chemical reactions as explained in Figure 2. This process is very slow since each step of the cycle may require a few tens of seconds. New technologies, such as enzymatic synthesis, are emerging. They are more environmentally friendly and promise lower costs. But the whole synthesis process remains slow.

Regardless of the technology, the key to increasing synthesis throughput and reducing costs lies in miniaturization and parallelization of the processes, as it has been done for sequencing 20 years ago. But the main challenge to solve, if semiconductor chips are targeted, is how to spatially separate chemical reactions on a flat surface without using physical containment.

The general idea, developed by different companies and research labs, is to implement a large array of electrochemical devices patterned on the top of semiconductor chips. Each device can be individually addressed via the semiconductor circuitry and will synthesize a specific pool of oligos. Schematically, a device can be considered as an area on the chip where thousands of identical oligos will be synthesized simultaneously. Depending on whether the devices are activated or not, chemical reactions in the device areas will have different behaviors, and, therefore, will be able to drive the synthesis of oligos.

In general, the entire synthesis process is performed step by step. To add a nucleotide, the chip is fed sequentially with the four A, T, G, C nucleotides. The electrochemical devices are activated when the desired nucleotide flows.

The final product is a library of hundreds of thousands oligos defined by the user at a small fraction of the cost of making each oligos individually with traditional synthesis techniques. Various electrochemical devices have been devised to control the chemical reactions and have been successfully implemented on CMOS chips. In the following we present two different techniques that illustrate how chemical reactions can be controlled.

The first is based on the local modification of the acid composition of the environment. The CustomArray company (Genscript) has developed a CMOS chip composed of an array of electrodes [6]. An electrode is a specific area where a pool of oligos can be synthetized. When the electrode is turned on, it creates a confined acid environment that deprotects oligos and allows the insertion of the next nucleotide. Figure 3 illustrates how the synthesis process work on a two-electrode system. The largest chip provided by CustomArray is composed of 92,918 electrodes with a diameter of 25 µm.

The CMOS chip developed by the Twist Bioscience company is based on the same electrical chemistry [3]. In their roadmap, three new generation of chips are envisaged, the last one having a synthesis capacity of 50 billions of oligos, representing 1 TBytes of data.

The second technique relies on regulating precisely the temperature of chemical reactions. It is currently developed by the Evonetix company [7]. Their approach uses a silicon chip, made by MEMS processing, that controls DNA synthesis at several thousand independently controlled reaction sites on the chip. By controlling the temperature of each reaction site, the growing strands of DNA are selectively deprotected, preparing them for new nucleotides to be added according to the intended DNA sequence.

But unlike the previous method, which generates oligos, double-stranded DNA can be produced directly by adding different thermally controlled steps after the oligo synthesis process. This technique also allows to obtain longer DNA molecules. It is also compatible with enzymatic DNA synthesis.

IV. SEQUENCING

From a DNA storage perspective, sequencing technologies are much more mature than synthetic technologies. A major breakthrough occurred in the mid-2000s with the massive parallelization and miniaturization of a sequencing method called "sequencing by synthesis" (SBS). The Illumina company is the leader in this technology. Today, its largest DNA sequencer can generate billions of high-quality sequences (<0.1% error rate) from the genomes of any living organisms.

Schematically, the SBS technology works as follows: to "read" a DNA molecule, it is first dehybridized (the

double strand is divided into its two complementary single strands) and one strand is kept as a template to reconstruct the original molecule. The synthesis (reconstruction of the molecule) is performed using modified nucleotides (dNTPs) with fluorescence capacity. Each time a new nucleotide is added to the double strand, a light treatment is performed to detect the type of nucleotide (A, C, G or T) based on its fluorescence characteristics.

Miniaturization and parallelization of SBS are primarily achieved using flow cell devices (glass slides). The last Illumina technology uses patterned flow cells with billions of nano wells at fixed location. One well is dedicated to deciphering a single DNA molecule. It is first amplified inside the nano well before to start the sequencing process itself. A scan of the entire flow cell is done each time a new nucleotide is inserted and the image is analyzed to deduce the type of nucleotide has been added.

Patterned flow cells and CMOS technologies can be coupled to speed-up data analysis as proposed by Illumina through its semiconductor sequencing technology. Each well is aligned to a CMOS photodiode that directly detects light emissions. Adding this possibility to the flow cells simplify greatly overall data analysis.

The SBS technology suffers from the fact that it can only produces short sequences, typically limited to a maximum of 300 nucleotides. Other technologies proposed by the Pacific Biosciences (PacBio) and Oxford Nanopore Technology (ONT) companies address this limitation. PacBio has developed a technology, called Single Molecule, Real-Time (SMRT) sequencing, which is also based on the detection of fluorescent nucleotides (dNTPs) during the synthesis of a template strand [5]. To our knowledge, no dedicated semiconductor technology is associated with this process.

ONT sequencing devices use flow cells that contain an array of thousands of small holes, called nanopores, embedded in an electro-resistant membrane. Each nanopore has its own electrode connected to a sensor chip capable to accurately measure the electric current flowing through the nanopore. When a molecule passes through a nanopore, the current is modified, producing a characteristic signature depending of the nucleotides present inside the nanopore. Typically, at a given time, five nucleotides are present inside the nanopore. The suite of signatures is then decoded to determine the composition of the DNA molecule (see Figure 4).

The largest sequencers (SBS technology) can currently read the equivalent of 6×10^{12} nucleotides in 24 hours, that is, theoretically 12×10^{12} bits. However, sequencing is a sort of random process that requires the prior constitution of libraries containing numerous instances of the molecules to be read. Only a fraction of the molecules will be sequenced. To be sure to read at least one instance of each molecule, one usually sequences 20 to 30 times more than necessary. Taking also into account the redundancy brought by the error correcting codes, we can estimate a throughput of about 0.5 MBytes/sec.

V. CONCLUSION

The storage of DNA is primarily aimed at the long-term preservation of "cold data". But for this solution to be viable, it is necessary to both greatly reduce costs and significantly increase the speed of writing (synthesis) and reading (sequencing). Today, the bottleneck is mainly synthesis. Optimizing this step requires massive miniaturization and parallelization. The solutions that are emerging to accelerate the synthesis process clearly indicate that the use of semiconductors is unavoidable.

ACKNOWLEDGMENT

This research was funded, in whole or in part, by the Labex CominLabs, Rennes and by the Inria dnarXiv exploratory action. A CC-BY public copyright license has been applied by the authors to the present document and will be applied to all subsequent versions up to the Author Accepted Manuscript arising from this submission, in accordance with the grant's open access conditions.

REFERENCES

[1] D. Reinsel, J. Gantz, and J. Rydning, "Data Age 2025", IDC white paper, April 2017.

[2] DNA Data Storage Alliance, "An introduction to DNA data storage", white paper, 2021

[3] A. Fernandez, From DNA Synthesis on Chips to DNA Data Storage, SDC 2021, September 28, 2021

[4] Illumina CMOS Chip and One-Channel SBS Chemistry, Technical Note 770-2013-054-B

[5] J. Eid and al., Real-Time DNA Sequencing from Single Polymerase Molecules, vol 323, Issue 5910, pp. 133-138, Science, 2009

[6] M. Caraballo, Oligo Pools: Design, Synthesis and Research Applications, Webinar, 2018

[7] S. Brooking, Putting DNA Synthesis in the hands of every researcher, Innovation in Pharmaceutical Technology, Nov. 2020

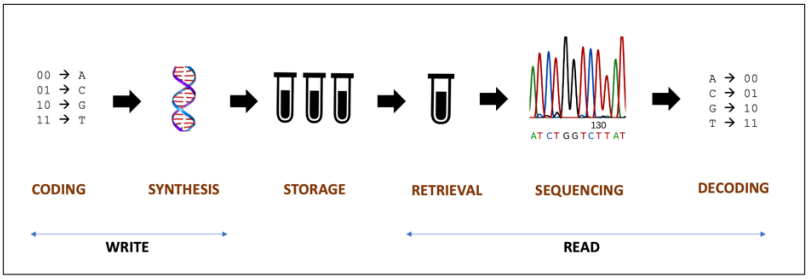

Fig. 1: DNA Storage pipeline. The whole DNA storage pipeline is split into 6 main steps. The two first steps correspond to the writing of data. The last three steps correspond to the reading of data.

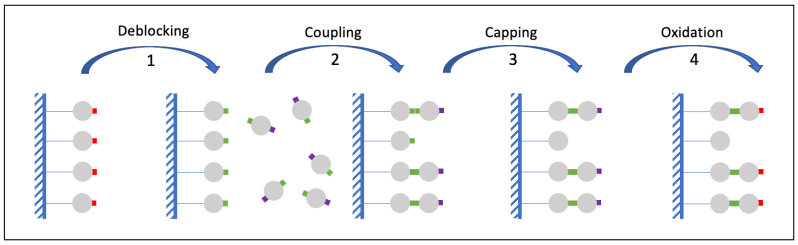

Fig. 2: Oligonucleotides synthesis process by phosphoramidite chemistry. The synthesis process adds nucleotides one by one, using a repeated 4-step cycle of deblocking, coupling, capping, and oxidation for each A, C, T, or G addition. **Deblocking**: this step remove the DMT protection group to allow the next nucleotide to be inserted. **Coupling**: A nucleotide is added to the oligonucleotide sequence. **Capping**: as the coupling is not 100% efficient, sometimes the coupling fails. Therefore uncoupled sequences could create errors in the synthesized molecule. To stop this, an unreactive group is added blocking further extension. **Oxidation**: required to stabilized nucleotide bounds.

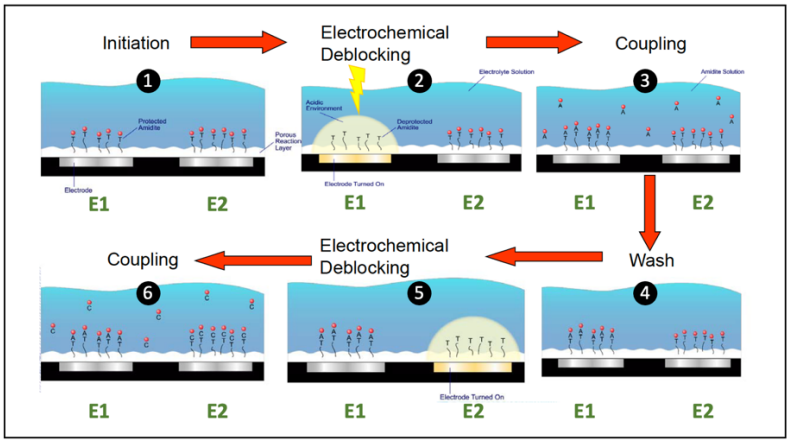

Fig. 3: CustomArray chip principle exemplified on a two electrode system. (1) At the beginning both electrodes house oligonucleotides starting with 'T'; (2) Electrode (E1) is turned on to allow insertion of a nucleotide; (3) 'A' nucleotides are flowed and can only couple with oligonucleotide of electrode E1; (4) free nucleotides are removed (wash operation) ; (5) Electrode E2 is turned on ; (6) 'C' nucleotides are flowed and can only couple with E2. Figure adapted from [6].

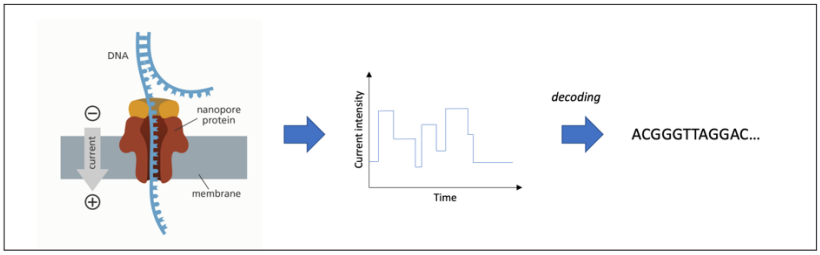

Fig. 4: ONT sequencing principle. One strand of the DNA molecule flows inside a nanopore protein attached to a membrane. As the DNA goes through the protein, intensity of the current, depending of the nucleotides inside the protein, is modified. Current intensity is measured and analyze to recover the sequence of nucleotides.